Communities and anonymization in graphs

Pierre Cazals

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris Dauphine

Communities and anonymization in graphs

Soutenue par
Pierre CAZALS
Le 10 décembre 2021

École doctorale n°543
Ecole doctorale SDOSE

Spécialité
Informatique

Composition du jury :

Eric Angel Professeur, Université d'Évry	*Rapporteur*
Cristina Bazgan Professeur, Université Paris Dauphine	*Directeur de thèse*
Cédric Bentz Maître de conférence HDR, CNAM	*Rapporteur*
Nadia Brauner Professeur, Université Grenoble Alpes	*Président du Jury*
Florent Foucaud Maître de conférence, Université Clermont Auvergne	*Examinateur*

MINES ParisTech | PSL

Contents

Introduction

The amount of collected data is increasing, whether in terms of the quantity of data, its variety or the number of actors collecting it. This growth is made possible by the ever-increasing accessibility to a computer terminal, to the Internet network and by the decrease in storage costs for servers. According to the research firm IDC, the volume of global storage would increase from 33 zettabytes in 2018 to 175 zettabytes in 2025. From the same source, 75% of this data is held by companies. The size of the European data market is estimated at €184 billion in 2020 and is expected to reach between €200 and 300 billion in 2025, according to the European Data Portal.

Data holders are finding it difficult to disclose this information without compromising the privacy of individuals. Through the GDPR (General Data Protection Regulation), we observe that a dynamic of tightening legislation around the holding of personal data is underway. Ultimately, in many cases, the survival of these databases will depend on the capacity of the holder to produce anonymous data to allow their exploitation without harming others.

Currently, data holders operate individually, ignoring the possibility of cross-checking with other data. A common practice called pseudonymization (encouraged by the GDPR) is to process data in such a way that it is impossible to attribute the data to an individual without the help of additional data. In most cases the remaining data can be used to re-identify individuals by matching the data with other existing databases. A known example is Latanya Sweeney's experience: She first showed that the combination of ZIP code, gender and birth date was unique for 87% of Americans [55]. In a second step, see Figure 1, she used two public databases, the "voter registration list for Cambridge Massachusetts" containing the quadruplet: Name, Address, ZIP code, Gender of the inhabitants of Cambridge and a second one, produced by the "Group Insurance Commission", responsible for purchasing insurance for the states employees, containing the medical records as well as the ZIP code, Gender and Birthday of each individual although by their Name. By a simple cross-check, she was able to identify the medical file of the governor of that state. Therefore, there is a challenge in ensuring sufficient anonymity while allowing data exploitation.

To contribute to this challenge, we propose to identify models that guarantee a "suitable" anonymization in graph databases. We will then seek to design algorithms that transform data in such a way that they respect these models while minimizing both the induced perturbations and the computation time. More precisely, our scope is to design efficient graph anonymization algorithms with performance guarantees and give an overview of the computational hardness. We try to focus on parameters as close as possible to practical situations, in term of anonymity level,

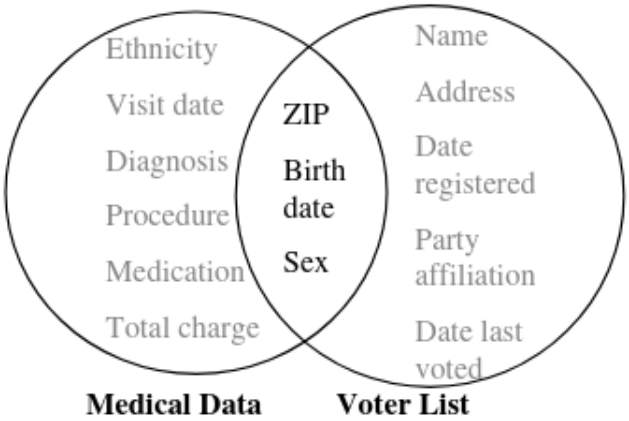


Figure 1: Data cross-checking

graph structure and data loss.

Furthermore we are interested in the structure of the graph database. Real networks are not random graphs: they present strong particularities and are generally sparse graphs with few high degree vertices. Moreover the distribution of edges is also locally inhomogeneous with strongly connected sets of vertices and very weak interaction between these groups (see example in Figure 2). This specific organization of real networks is called community partition [30].

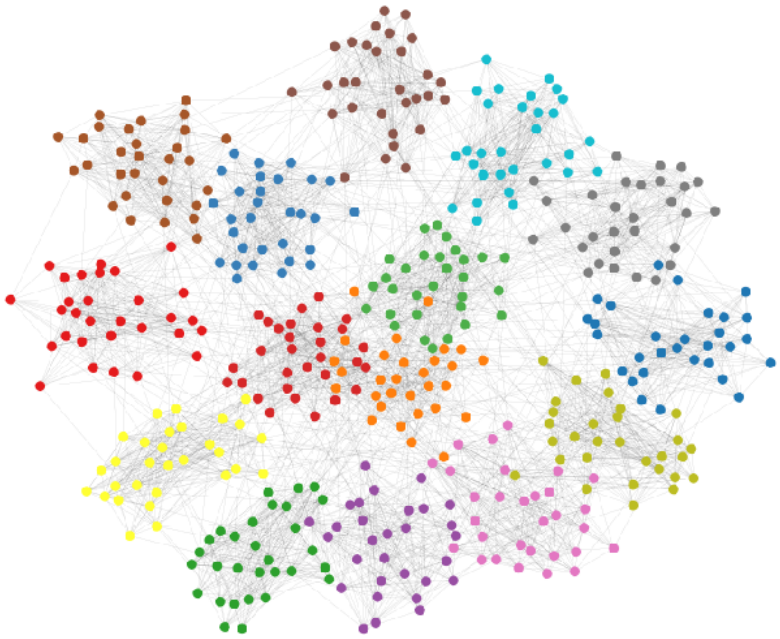


Figure 2: Graph with highlighted communities

More concretely, societies present a wide variety of social groups: relational (family, work, friendship), geographical (village, city, country), cultural (language,

hobbies, music), etc.. On the other hand, the internet has allowed the creation of communities that did not exist before (forums, social networks, video games).

Although the notion of community may seem intuitive, giving a formal definition is more difficult. Indeed, there is no universal definition and often the definition will depend on the intended application. However, there is a common point between the different definitions of communities: we wish to have the maximum number of edges inside the communities and therefore the minimum outside. We propose to contribute to this field by studying a dense subgraph partition problem, where the density represents the cohesivity of each part, from the perspective of computational complexity [21].

This document is organized in three chapters. The first one presents the background required to understand the material presented in this thesis. We first present central notions related to anonymization, anonymization models, privacy, data utility, attacks; then notions on community partition and finally we define notations related to graph theory used in this document. The second chapter focuses on a classic graph anonymization model. We investigate the existence of a solution as well as the NP-hardness with chosen graph modification. We also design an approximation algorithm along with a simple exact algorithm. In a third chapter, we study a problem of partitioning into dense subgraphs. We classify the difficulty of the problem for several classes of graphs, including one relatively similar to social network graphs (sparse). We then design an efficient approximation scheme for dense graphs.

Chapter 1

Preliminaries

In this section we will first define the notations used and then we present the general key notions of the subjects studied in this manuscript, namely anonymization and community partition.

1.1 Notations

Sets

A set is a collection of distinct elements without order. Let A and B be two sets and F a set of sets.

$x \in A$	x is an element of A	
$x \notin A$	x is not an element of A	
$\lvert A \rvert$	Number of elements of A	
$A \cap B$	intersection of A and B	$\{x \mid x \in A \wedge x \in B\}$
$A \cup B$	union of A and B	$\{x \mid x \in A \vee x \in B\}$
$A \setminus B$	A minus B	$\{x \mid x \in A \wedge x \notin B\}$
$\bigcup_{V \in F} V$	union of F sets	$\{x \mid \exists V \in F,\ x \in V\}$
$\bigcap_{V \in F} V$	intersection of F sets	$\{x \mid \forall V \in F,\ x \in V\}$
$\mathcal{P}(A)$	set of all partitions of A	$\{\mathcal{P} \mid \bigcup_{V \in \mathcal{P}} = A \wedge \forall U, W \in \mathcal{P},\ U \cap W = \emptyset\}$
$A \subseteq B$	A is a subset of B	$\{x \mid x \in A \wedge x \notin B\} = \emptyset$
$A \nsubseteq B$	A is not a subset of B	$\{x \mid x \in A \wedge x \notin B\} \neq \emptyset$

Miscellaneous

Let x be a real number. Let k and n be two integers.

(x, k)	couple of x and k	ordered pair of elements
$\lfloor x \rfloor$	Floor value of x	$\lfloor x \rfloor = \max\{a \in \mathbb{Z}, a \leq x\}$
$\binom{n}{k}$	Binomial coefficient	$\frac{n!}{(n-k)!k!}$

Graphs

Let $G = (V, E)$ be a couple such that V is an arbitrary set and $E \subseteq \{\{u, v\} \mid \{u, v\} \subseteq V \wedge u \neq v\}$. G is an undirected and unweighted graph without loops and multiple edges. In this document, we assume that all graphs are of this shape. Let X and $\{u, v\}$ be subsets of V and v an element of V.

V	the vertex set	arbitrary set
E	the edge set	$E \subseteq \{\{u, v\} \mid \{u, v\} \subseteq V \wedge u \neq v\}$
n_G	the number of vertices	$n_G = \lvert V \rvert$
m_G	the number of edges	$m_G = \lvert E \rvert$
$\mathcal{N}_G(v)$	the neighbourhood of v	$\mathcal{N}_G(v) = \{u \mid u \in V \wedge \{u, v\} \in E\}$
$d_G(v)$	degree of v	$d_G(v) = \lvert \mathcal{N}_G(v) \rvert$
$inc_G(v)$	the set of edges incident to v	$inc_G(v) = \{\{u, v\} \mid \{u, v\} \in E\}$
$G[X]$	the graph induced by X	$G[X] = (X, \{\{u, v\} \mid \{u, v\} \in E \wedge \{u, v\} \subseteq X\})$
$\overline{G}$	the complementary graph	$\overline{G} = (V, \{\{u, v\} \mid \{u, v\} \subseteq V \wedge u \neq v \wedge \{\{u, v\} \notin E\})$
Δ_G	maximum degree	$\max\{d_G(u),\ u \in V\}$
$D_G(x)$	Vertices of degree x	$\{u \in V \mid d_G(u) = x\}$
$D_G^{\delta}(x)$	Vertices of degree $\geq x$	$\{u \in V \mid d_G(u) \geq x\}$

The edge $\{u, v\}$ is denoted uv and if the underlying graph G is clear from the context, we omit the subscript G.

Graph classes

$\mathcal{G}$	graphs	$\{G = (V, E) \mid E \subseteq \{uv \mid \{u, v\} \subseteq V \wedge u \neq v\}\}$
$\mathcal{G}_c$	connected graphs	$\{G \in \mathcal{G} \mid \forall \{A, B\} \in \mathcal{P}(V),\ \exists u \in A,\ \exists v \in B,\ uv \in E\}$
r-REG	r-regular graphs	$\{G \in \mathcal{G} \mid \forall v \in V,\ d_G(v) = r\}$
CUB	cubic graphs	3-regular graph
SG	stable graphs	0-regular graph
BP	bipartite graphs	$\{G \in \mathcal{G} \mid \exists \{A, B\} \in \mathcal{P}(V),\ \{G[A], G[B]\} \subseteq SG\}$
3-C	3-colorable graphs	$\{G \in \mathcal{G} \mid \exists \{A, B, C\} \in \mathcal{P}(V),\ \{G[A], G[B], G[C]\} \subseteq SG\}$
TR	trees	$\{G \in \mathcal{G} \mid m_G = n_G - 1 \wedge G \in \mathcal{G}_c\}$
CAT	caterpillars	$\{G \in TR \mid \forall u \in V,\ \lvert \mathcal{N}(u) \cap D_G^{\delta}(2) \rvert \leq 2\}$
$\mathcal{G}(n, m)$	(n, m)-graphs	$\{G \in \mathcal{G} \mid \lvert V \rvert = n \wedge \lvert E \rvert = m\}$

Graph modifications

Let $G = (V, E)$ be a graph. Let $\{u, v, u', v'\} \subseteq V$ and $w \notin V$ such that $uv \in E$ and $\{u'v', uv'\} \cap E = \emptyset$:

V-del	vertex deletion	$G' = (V \setminus \{u\}, E \setminus inc_G(u))$
V-add	vertex addition	$G' = (V \cup \{w\}, E \cup E'),\ E' \subseteq \{wv \mid v \in V\}$
E-del	edge deletion	$G' = (V, E \setminus \{uv\})$
E-add	edge addition	$G' = (V, E \cup \{u'v'\})$
E-mov	edge move	$G' = (V, E \setminus \{uv\} \cup \{u'v'\})$
E-rot	edge rotation	$G' = (V, E \setminus \{uv\} \cup \{uv'\})$

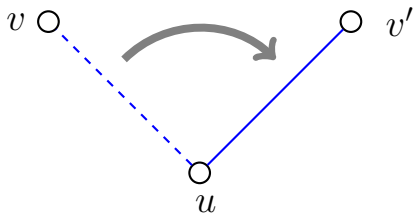


Figure 1.1: An **edge rotation** (uv, uv') from uv to uv'

Definition 1.1.1. *Let G and G' two graphs of $\mathcal{G}(n, m)$. The **edge rotation distance** between G and G' is the minimum number of edge rotations required to transform G into G'. Note that it is symmetric.*

Degree sequence

Definition 1.1.2. *Given a graph $G = (V, E)$ of order n, **the degree sequence** S_G of G is the non-increasing sequence of its vertex degrees, $S_G = (deg(v_1), \ldots, deg(v_n))$, $deg(v_1) \geq deg(v_2) \geq \cdots \geq deg(v_n)$.*

Definition 1.1.3. *A sequence D of non-negative integers $D = (d_1, d_2, \ldots, d_n)$ is **graphic** if there exists a graph G such that its degree sequence coincides with D.*

As follows from Erdős-Gallai theorem (see e.g. [24]) the necessary and sufficient conditions for a non-increasing sequence $D = (d_1, d_2, \ldots, d_n)$ to be graphic are:

$$\sum_{i=1}^{n} d_i \text{ is even} \tag{1.1}$$

$$\sum_{i=1}^{\ell} d_i \leq \ell(\ell - 1) + \sum_{i=\ell+1}^{n} \min(d_i, \ell) \text{ holds for any } 1 \leq \ell \leq n. \tag{1.2}$$

1.2 Database, Attack and Anonymization model

A relational database is a database in which information is presented in a tabular form, usually in two dimensions. The model was introduced by Edgar Codd in 1970, [17]. According to this model, a database consists of one or more relations. The rows of these relations are called tuples. The columns are called attributes.

Table 1.1 is an example of a relational database, with different kinds of attributes. Identifiers are attributes that allow to associate a tuple of attributes to an individual;

	Identifier	Non-sensitive			Sensitive
	Name	Gender	Birthdate	ZIP code	Disease
111	Michel	Male	1944	34300	Cancer
112	Ginette	Female	1944	34700	Cancer
113	Mireille	Female	1944	34000	Cancer
114	Umeko	Male	1942	34000	Viral infection
115	Nina	Female	1942	34300	Heart Disease
116	Conchita	Female	1942	34700	Heart Disease
117	Rolande	Female	1947	34300	Back pain
118	Olivier	Male	1947	34000	Vomiting
119	Scarlett	Female	1947	34700	Obesity

Table 1.1: An extract of a medical relational database (people born between 1942 and 1947 that have 34*** as ZIP code)

they are unique. Attributes, other than identifiers, that the patient does not want to be disclosed are sensitive attributes. Other attributes are non-sensitive.

Let us suppose that a researcher wants to use the database from Table 1.1. However, he does not have the right to dispose of the information contained inside because it is too private. For this reason we want to transform the database so that personal information cannot be linked to its owner.

To overcome this problem, there are different models of anonymization that allow for some use of the data while guaranteeing a level of anonymity. Anonymity should not be seen as a binary value (anonymous or not) in the sense that the only transformation that makes the data completely anonymous is the complete deletion of the data. It is the same for the data utility, it is more of a gray area. From this observation, a good anonymization algorithm will therefore be an algorithm able of offering a good compromise between destroying the "most identifying" information and keeping the "most relevant" for a specific purpose.

We assume that a potential malicious adversary searching for information about a particular person knows whether that person belongs to the database or not. Moreover, in any anonymization process, identifiers are removed.

Since most models were designed to solve issues caused by specific attacks, we will first present attacks and models following a historical order. Then, we will define the data utility.

More formally, the main goal of anonymization is to protect the database against the following issues:

Definition 1.2.1. *There is an **identity disclosure** when observations of the anonymized data allows the adversary to deduce the identity of the subject in the data set. After this, the adversary can link all the sensitive attribute values in the record to the subject's identity, which clearly violates the subject's privacy.*

Definition 1.2.2. *There is an **attribute disclosure** when the adversary can infer some sensitive information about an individual without identifying individual's record in the published data set. The adversary can link one sensitive attribute to the subject's identity, which also violates the subject's privacy.*

	Identifier	Non-sensitive			Sensitive
	Name	Gender	Birthdate	ZIP code	Disease
111	Michel (removed)	Male	1944	34300	Cancer
112	Ginette (removed)	Female	1944	34700	Cancer
113	Mireille (removed)	Female	1944	34000	Cancer
114	Umeko (removed)	Male	1942	34000	Viral infection
115	Nina (removed)	Female	1942	34300	Heart Disease
116	Conchita (removed)	Female	1942	34700	Heart Disease
117	Rolande (removed)	Female	1947	34300	Back pain
118	Olivier (removed)	Male	1947	34000	Vomiting
119	Scarlett (removed)	Female	1947	34700	Obesity

Table 1.2: The extract with the identifiers removed

	Identifier	Non-sensitive		
	Name	Gender	Birthdate	ZIP code
281	Michel	Male	1944	34300
292	Ginnette	Female	1944	34700
315	Mireille	Female	1944	34000
362	Umeko	Male	1942	34000
389	Nina	Female	1942	34300
401	Conchita	Female	1942	34700
403	Rolande	Female	1947	34300
452	Olivier	Male	1947	34000
499	Scarlett	Female	1947	34700

Table 1.3: An extract of the voter list (public data)

The matching shared attribute attack is an example of identity disclosure. The homogeneity attack is an example of attribute disclosure.

The first anonymization model consisted solely of deleting identifiers. The following attack is the same as the one presented in the introduction.

Attack: Matching shared attribute

This attack consists in re-identifying a database where names have been deleted using another non-anonymous database that shares some non-sensitive attributes. Table 1.2 is the database of Table 1.1 where identifiers were removed. Table 1.3 is a voter registration list extract, a public database accessible to all. We can observe that there is only one man born in 1944 with the postal code 34300. We deduce that this man is Michel and that he has cancer.

Shared non-sensitive attributes used to re-identify records leads to the following definition :

Definition 1.2.3. *A **quasi-identifier** is a chosen combination of non-sensitive attributes such that the combination is often unique and allows the target to be re-*

	Identifier	Non-sensitive			Sensitive
	Name	Gender	Birthdate	ZIP code	Disease
111	Michel (removed)	Unspecified	1944	34***	Cancer
112	Ginette (removed)	Unspecified	1944	34***	Cancer
113	Mireille (removed)	Unspecified	1944	34***	Cancer
114	Umeko (removed)	Unspecified	1942	34***	Viral infection
115	Nina (removed)	Unspecified	1942	34***	Heart Disease
116	Conchita (removed)	Unspecified	1942	34***	Heart Disease
117	Rolande (removed)	Unspecified	1947	34***	Back pain
118	Olivier (removed)	Unspecified	1947	34***	Vomiting
119	Scarlett (removed)	Unspecified	1947	34***	Obesity

Table 1.4: 3-anonymization of the extract

identified.

In the Latanya Sweeney's example [56], see Figure 1, quasi-identifiers were the triplets address, ZIP code and gender used to link the two databases.

For example, the following set of information : web browser version, add-ons, system version, language, screen size, etc. that everyone gives to websites is unique for most people, then it can be considered as a quasi-identifier. This quasi-identifier can be used to link your activities on different websites. If a user wants to know how unique he is, several websites allow to do it. *AmIUnique* is the most known. Methods to improve your anonymity on web browsing can be found on the Wikipedia page *Device Fingerprint*.

Model: k-anonymity

After studying the problem of re-identification, Sweeney proposes an anonymization model to overcome this flaw, the k-anonymization [56]. It is based on the notion of quasi-identifiers presented above.

The idea is to ensure that there is no unique quasi-identifier. The model requires that there are at least k occurrences of each combination of quasi-identifier that appears in the database.

Definition 1.2.4. *A database is said to be k-**anonymous** if each of the possible combinations of quasi-identifiers corresponds to 0 or at least k individuals.*

To k-anonymize a table, there are several possibilities: **generalize**, **delete**, **add** or **modify** data.

Coming back to the example of Table 1.1, if we want to 3-anonymize it, we can modify some non-sensitive attributes of the quasi-identifier used before. For instance, it is possible to not display the gender and instead of giving the ZIP code, simply give the department number. In all cases, we still remove the identifiers. Table 1.4 is the resulting 3-anonymized table. The statistical relevance of the data will change very little and the previous attack is no longer possible.

Attack: Homogeneity

This attack, designed by Machanavajjhala et al. [48], leverages the case where all values of a sensitive attribute of individuals having the same quasi-identifier are also equal. Going back to the example in Table 1.4, we can observe that all people born in 1944 with postal code 34*** have cancer. Without knowing precisely which line Michel corresponds to, we can deduce that he has cancer. It's an attribute disclosure, in any case we can not guess Michel's rows.

This problem is not as unlikely as it seems. Suppose that we have one sensitive attribute that can take 3 values in an equiprobable way. If five entries share the same quasi-identifier, there is 3 in 3^5 chances, i.e. $\frac{1}{81}$, that they share also the same sensitive attribute. On tables with hundreds of thousands of entries, it is almost sure to encounter attribute disclosures.

Attack: Background knowledge

Here, it is assumed that the adversary has background knowledge on its target. We adapt an example given by Machanavajjhala et al. [48]. Alice has a pen-friend named Umeko who is admitted to the same hospital as Michel and whose patient records also appear in the table shown in Table 1.4. Alice knows that Umeko is a Japanese man born in 1942 in a department of ZIP code 34000. Based on this information, Alice learns that Umeko's information is contained in record number 114, 115 or 116 (see Table 1.4). Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection. In this particular example it is an identity disclosure because there is only one person with a viral infection. In other cases it would just be an attribute disclosure.

Model: ℓ-diversity

The ℓ-diversity model, also designed by Machanavajjhala et al. [48], is a k-anonymity patch which aims to prevent homogeneity and background knowledge attacks. The model is based on two notions, that of equivalence classes formed by quasi-identifiers and diversity.

Definition 1.2.5. *For a given set of quasi-identifiers Q in a database, each row that contains Q belongs to the* ***equivalence class*** *of Q.*

In the example of Table 1.4, rows 111, 112 and 113 belong to the same equivalence class because they share the same quasi-identifier : 1944, 34***.

There are several notions of diversity. Here is one of them:

Definition 1.2.6. ***Distinct ℓ-diversity*** *ensures that at least ℓ distinct values for the sensitive attribute in each equivalence class exists. A table satisfies ℓ-diversity if all its equivalence classes have ℓ-diversity.*

	Identifier	Non-sensitive			Sensitive
	Name	Gender	Birthdate	Zip code	Disease
114	Umeko (removed)	Unspecified	194*	34000	Viral infection
118	Olivier (removed)	Unspecified	194*	34000	Vomiting
113	Mireille (removed)	Unspecified	194*	34000	Cancer
111	Michel (removed)	Unspecified	194*	34300	Cancer
115	Nina (removed)	Unspecified	194*	34300	Heart Disease
117	Rolande (removed)	Unspecified	194*	34300	Back pain
116	Conchita (removed)	Unspecified	194*	34700	Heart Disease
112	Ginette (removed)	Unspecified	194*	34700	Cancer
119	Scarlett (removed)	Unspecified	194*	34700	Obesity

Table 1.5: A 3-anonymization, 3-diversification of the database

It is to be noted that it is necessary to perform k-anonymization before ℓ-diversification while making sure that $k \geq \ell$. Otherwise, we would run the risk that some classes are smaller than ℓ. In real life settings, k is usually set much higher than ℓ.

By generalizing the birthdate instead of the ZIP code we obtain a 3-anonymous, 3-diverse table. Note that it is not always easy, sometimes it involves more data loss. In Table 1.5, order of rows has been swapped to highlight the classes.

This process makes it possible to withstand homogeneity and background knowledge attacks. The ℓ-diversity reduces the chances to identify a specific attribute to $\frac{1}{\ell}$ with the homogeneity attacks. For the background knowledge attack, if there are ℓ possible diseases, Alice will need a lot of information to eliminate the ℓ minus one wrong ones.

We can observe that not all cases can be well handled by this method since sensitive attributes in real life are not necessarily diverse. Sometimes, we will be forced to modify the sensitive attributes. For example, if we consider a database where the sensitive attribute is whether a patient has contracted HIV or not, statistically (in France) there would be more than $\frac{1}{500}$ people who have HIV. Dealing with this consideration, if we want to have a 50-anonymous and 2-diverse table, we would have to add around 9 false positives for a real one.

Attack: Complementary release

In this attack, the opponent will use two different k-anonymous ℓ-diverse releases from the same database. Consider the 3-anonymous, 3-diverse table of Table 1.5 and another 3-anonymous, 3-diverse table of the same database in Table 1.6. The opponent knows that Michel is a man born in 1944 with a postal code of 34300 Table 1.3. Using Table 1.5, we learn that a person born in 194* with ZIP code 34300 can have either cancer, back pain or heart disease. In the second table, male individuals born in 194* with ZIP code 34*** can have either cancer, vomiting or viral infection. Michel belongs to both of these, we deduce that Michel has cancer.

	Identifier	Non-sensitive			Sensitive
	Name	Gender	Birthdate	Zip code	Disease
111	Michel (removed)	Male	194X	34***	Cancer
118	Olivier (removed)	Male	194X	34***	Vomiting
114	Umeko (removed)	Male	194X	34***	Viral infection
115	Nina (removed)	Female	194X	34***	Heart Disease
113	Mireille (removed)	Female	194X	34***	Cancer
116	Conchita (removed)	Female	194X	34***	Heart Disease
117	Rolande (removed)	Female	194X	34***	Back pain
112	Ginette (removed)	Female	194X	34***	Cancer
119	Scarlett (removed)	Female	194X	34***	Obesity

Table 1.6: An other 3-anonymization, 3-diversification of the database extract

Model: t-closeness

t-closeness [42] is an extension of ℓ-diversity to solve the previous attack. It explores a statistical method, mainstream now, to produce anonymity. As it is not in the scope of this thesis, we direct the reader to [42] for further details.

The idea of t-closeness is to control the difference between the distribution of sensitive attributes in the same class and in the whole table.

Definition 1.2.7. *An equivalence class is said to satisfy* t***-closeness*** *if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table does not exceed a threshold t. A table satisfy t-closeness if all equivalence classes satisfy t-closeness.*

Guaranteeing this property solves the previous problem since each equivalence class is t-close (very similar) to the whole data. The downside is that data can be seriously harmed, first, by applying the mandatory k-anonymization process then with the t-closeness one.

Giving a representative example would require a too large table.

Model: Differential Privacy

More recently, Dwork et al. [23] proposed a new model of anonymization, called differential privacy. As for t-closeness, differential privacy guarantees are based on statistical properties. The point of view is a bit different though. The database cannot be accessed directly, only through a proxy algorithm whose output satisfies anonymization requirements. The algorithm performs a probabilistic data processing and the user only gets the output. The algorithm guarantees that for two similar inputs (inquiries about the data), the returned results are similar. More formally:

Definition 1.2.8. *Let ϵ be a positive real number and A a probabilistic algorithm whose input is a dataset. Let $im(A)$ be the image of A. A is an ϵ-differentially private algorithm if, for every D_1 and D_2 that differ only on one element and for every $S \subseteq im(A)$: $P[A(D_1) \in S] \leq e^{\epsilon} \times P[A(D_2) \in S]$.*

Some models consider a bias parameter δ.

The method yields a powerful anonymization although dependent on the opponent knowledge. Even if the latter knows that his target has a property $\mathscr{P}$ with a probability p, he can not increase his certainty by more than $f(\epsilon, p)$. To reach this anonymity guaranty, noise is needed, the challenge being finding a distribution of this noise that does not damage excessively the data while ensuring the ϵ-differential privacy.

Model: k-anonymity and Differential privacy

There are several papers about differentially private algorithms of k-anonymity. We can cite the work of Wang et al. [58] or Sorias-Coma [54]. The goal is to produce differentially private k-anonymity algorithms with a higher ϵ as usual to limit data damage while keeping the advantage of k-anonymization.

Synthesis

Although the models presented above offer increasing guarantees of anonymity, they also require further modification of the data to achieve them. For example, to achieve ℓ-diversity, you have to k-anonymize and then re-process the sensitive attributes. There is therefore no model that strictly dominates another and if we are even aware of the possible attacks, we will take care to choose the simplest model that resists these attacks.

Since k-anonymization is a necessary independent first step for many models, we believe it is important to help provide a better understanding of the mechanisms around data transformation in order to satisfy the model. We will focus on the existing translations of the model in the graph database universe.

1.3 Extension of k-anonymity to graphical databases

A graph database is a database stored in a graph structure. The entities are generally represented by vertices and their attributes by the vertex label. In addition one can add a link between two entities, represented by an edge. This link can also have attributes, stored in the label of the edge. Such databases are also called social network database, since most of social networks are stored in a graph database.

In 2008, Liu and Terzi proposed an adaptation of the classical k-anonymization on relational network for the context of graph database, they named it *k-degree-anonymization* [44]. This model is designed for the framework of graph database, especially for social network graphs. Their idea is to identify the possible breaches in this context and propose a dedicated solution. They grouped privacy breaches in social networks in three categories : 1) *identity disclosure*: the identity of the individual who is associated with the node is revealed, this is the same definition as for relational database; 2) *link disclosure*: sensitive relationship between two individuals is disclosed (the opponent knows the existence of an edge); and 3) *content*

disclosure: the privacy of the data associated with each node is breached (i.e., label content). If the rows of the database are not independent, i.e., there is an attribute that takes as value an other row, a link disclosure in graph database can be seen as a disclosure of this kind of attribute in a relational database. If such an attribute does not exist, link disclosure is meaningless in relational databases. Content disclosure can be associated with attribute disclosure. k-degree-anonymity focuses on the identity disclosure, which does not imply anything for the other disclosures (as we see in the homogeneity attack for example). To handle content disclosure, they referred to known anonymization methods, standard privacy-preserving data mining techniques [1]. To preserve against link disclosure, they are based on link mining robustness techniques [29, 61].

The formal definition of a k-degree-anonymous graph is:

Definition 1.3.1. *A sequence of integers $D = (d_1, d_2, \ldots, d_n)$ is called **k-anonymous** where $k \in \{1, \ldots, n\}$, if for each element d_i from D there are at least $k-1$ other elements in D with the same value. A graph G is called **k-degree-anonymous** if its degree sequence is **k-anonymous**. The vertices of the same degree correspond to a **degree class**.*

1.4 Community partitioning

The literature around communities is wide and has been gathered and synthesized by S. Fortunato in [26].

In graphs representing human interactions, one of the most relevant features is the identification of social groups. Intuitively, a social group will appear as a set of vertices that are very connected to each other and not very connected to others. Often these clusters can be considered as independent components of the graph interacting with each other.

Social networking communities

Networks representing the social links that unite individuals have long been studied by sociologists and computer scientists [53, 59]. The emergence of new telecommunication technology has offered new modes of interaction between individuals as well as a facility of external observations for the researchers. The best known examples are the telephone network and the Internet. In these social networks it is easy to observe the interactions of millions of individuals. In this context, communities can take the form of friendship or family circle, people sharing common interests etc.

In 2008, Blondel et al. [10] analyzed the network formed by the interactions of different users of a Belgian telephone operator. The graph contains 2.6 million vertices and edges are weighted by the cumulative communication time between the two adjacent vertices (see Figure 1.2). As said before the precise definition of community remains a bias of the authors but it shares some common properties. Indeed, they succeeded in extracting 261 groups of more than one hundred vertices which are separated into two specific groups: French and Dutch speakers. This is

of course completely unlikely in a uniformly drawn graph. We can see that a social network which reflects connection of human society is far from being a random graph.

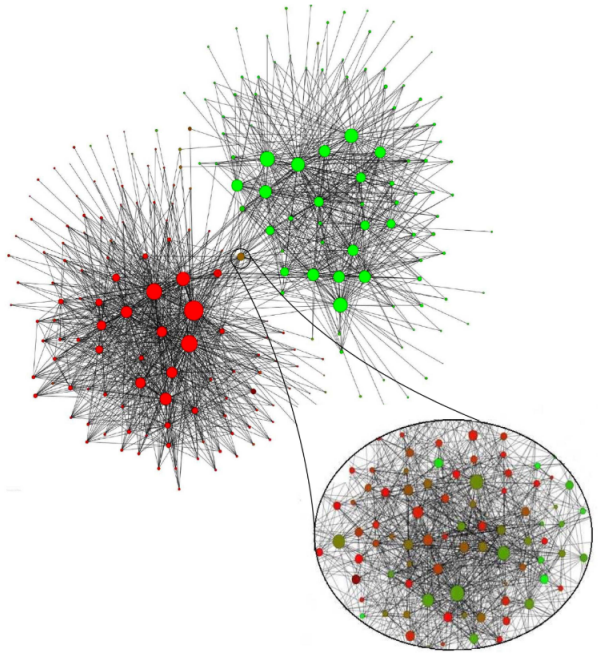


Figure 1.2: Belgian telecommunication : Two languages, two worlds and very few connections.

Red et al. [51] used Facebook data to reconstruct a relational network between students for different American universities (see Figure 1.3). The objective was to study the link between online and offline life. They found that communities were structured around graduating classes or dormitories depending on the university (and the presence of dormitories).

A first definition of community: the clique and its relaxations

A first, somewhat restrictive, definition by Luce and Perry [47] stipulates that members of a community are related to all the others. However, we can see that this definition is extremely restrictive: a subgroup of a thousand people interconnected except a few ones seems to be a cohesive group but is not a community with this definition. Moreover, it was found that communities are not symmetrical groups of individuals but a hierarchy in the sense that there are highly connected core vertices and loosely connected peripheral vertices within the same community [53, 59]. Finally, the problem of finding a clique of maximum size belong to Richard Karp's original 21 problems, as well as partitioning a graph into a minimum number of cliques [39] (see Figure 1.4). Some works have been proposed based on a relaxation of the notion of clique, called t-clique by sociologists [46, 2] (i.e. subset of vertices of maximum diameter t in G, shortest paths can be made up in part of vertices outside the t-clique.). The notion of k-club (i.e. subgraph with maximum diameter k) can overcome this weakness but it is NP-hard to compute [34].

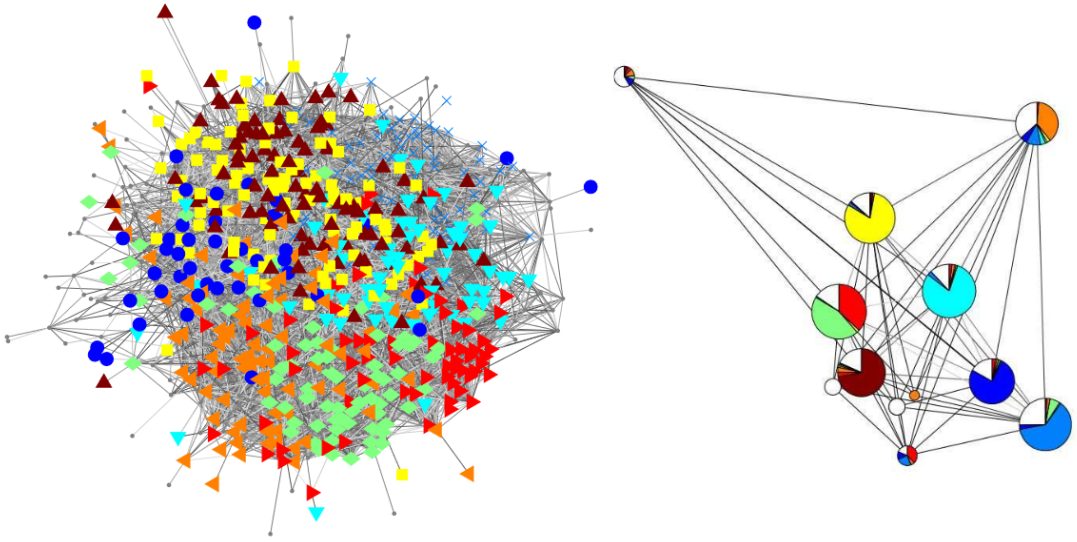


Figure 1.3: Caltech students' facebook friends network, colors/shape indicate the dormitories (Left). Community visualization with respect to the dormitories affiliation (Right). Pie charts represent the proportion of the number of links to each community, including itself.

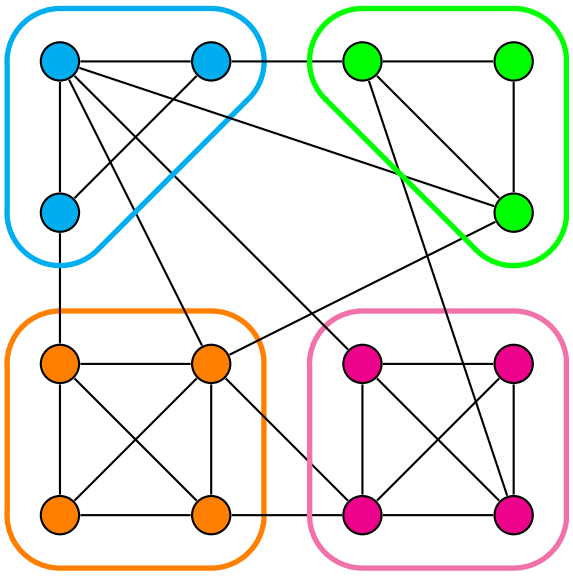


Figure 1.4: Partition into clique

Formalization of the notion of community: quality function

In this part the community is defined around the notion "there are more edges inside than outside". However this definition is quite open and we can find a lot of definitions that satisfy this postulate. On the other hand, the whole graph systematically satisfies this property, so we have to find something to optimize the size of the communities found. The most studied solution consists in seeing the communities as the output of a given algorithm, without prior definition. Most of these

algorithms rely on the notion of internal and external degree. Consider P a partition of the vertices of a graph G. For any vertex v of a part C of P we define $v_{int} = |\{u \mid u \in C \wedge u \in \mathcal{N}(v)\}|$ and $v_{ext} = |\{u \mid u \notin C \wedge u \in \mathcal{N}(v)\}|$. In a similar way we define $C_{int} = \frac{\sum_{v \in C} v_{int}}{|C| \times (|C|-1)}$ and $C_{ext} = \frac{\sum_{v \in C} v_{ext}}{|C| \times (n-|C|)}$ (see Figure 1.5).

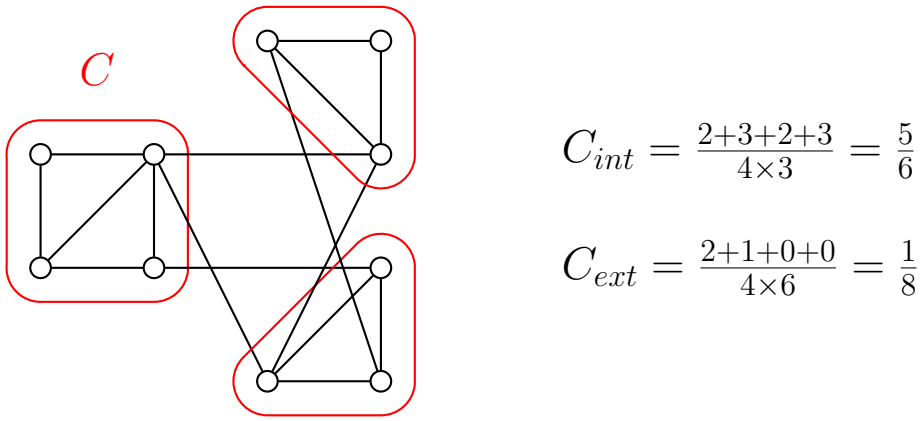


Figure 1.5: Internal and External connections of a community C.

Intuitively, a good community would be a subset C of vertices where C_{int} is much larger than C_{ext}. Finding a good trade-off between these two quantities is the goal of most community detection algorithms. For example, Mancoridis et al. tried to find a partition that maximizes the sum of the differences $C_{int} - C_{ext}$ of each part in [49].

On the other hand, another necessary natural condition is the connectivity of the communities. It is often induced by the trade-off optimization but sometimes it must be imposed as a constraint. Using this basic requirement many community definitions have been proposed depending on the application.

Partitioning into communities using quality function

Historically, researchers have mainly focused on cutting a graph into a predefined number of clusters (semi conductor design [40], parallel processing [50] and unsupervised learning [45] (k-means)). In the context of communities, we generally do not know the number of communities in advance and imposing it could provide undesirable results, such as the merging or splitting of cohesive groups.

We are therefore interested in providing algorithms that split the graph into an undefined number of parts. Although it is easy to split a graph into cohesive communities, not all partitions are interesting. In order to compare different partitions, we can use a criterion that ranks them between each other. Often this criterion is materialized by what is called a quality function, i.e., a function that takes as input a graph partition and returns a value in this context. Basically the partitions with high score are better than partitions with low score, so we will naturally look for a partition that maximizes it. However, it is important to keep in mind that the choice of the criterion greatly influences the shape of the partitions found and will therefore be related to the desired application.

Chapter 2

Edge rotations for degree anonymization

In this chapter, we investigate the computational complexity of Min Anonymous-Edge-Rotation defined as follows: an input to the problem is an undirected graph $G = (V, E)$ with n vertices and m edges and an integer $k \leq n$. The goal is to find a shortest sequence of edge rotations that transforms G into a *k-degree-anonymous* graph, if such a sequence exists. Recall that a k-degree-anonymous graph is a graph such that its degree sequence is k-anonymous (see Section 1.1). First, we will define the notions used in this chapter and review the state of the art of k-degree-anonymization.

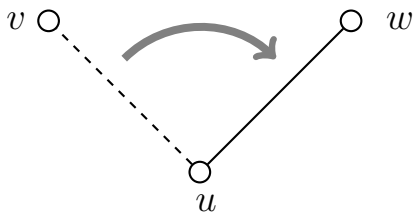


Figure 2.1: An edge rotation (uv, uw) from uv to uw

2.1 Preliminaries

In this chapter we assume that all graphs are undirected, without loops and multiple edges, and not necessary connected graphs. Let $\mathcal{G}(n, m)$ be the set of all graphs with n vertices and m edges.

Definition 2.1.1. *Let $G, G' \in \mathcal{G}(n, m)$. We say that G' can be obtained from G by an **edge rotation** (uv, uw) if $V(G) = V(G')$ and there exist three distinct vertices u, v and w in G such that $uv \in E(G)$, $uw \notin E(G)$, and $E(G') = (E(G) \setminus \{uv\}) \cup \{uw\}$, see Figure 2.1.*

The results of this section were published in [6, 7].

Remark 2.1.1. Let G be a graph. For the vertices u, v, w in G the edge rotation (uv, uw) modifies G into the graph G' such that $d_{G'}(v) = d_G(v) - 1$, $d_{G'}(w) = d_G(w)+1$, and the degree of the other vertices is not changed. Let define a $(+1, -1)$-degree modification of the degree sequence $D = (d_1, \ldots, d_n)$ in such a way that $d_i := d_i + 1$, $d_j := d_j - 1$ for any two indices i, j such that $i, j \in \{1, \ldots, n\}$.

In this chapter we study the following anonymization problem:

MIN ANONYMOUS-EDGE-ROTATION
Input: (G, k) where $G = (V, E)$ is an undirected graph and k a positive integer, $k \in \{1, \ldots, |V|\}$.
Output: If there is a solution, find a sequence of a minimum number $\ell + 1$ of graphs $G_0 = G, G_1, G_2, \ldots, G_\ell$ such that G_{i+1} can be obtained from G_i by one edge rotation, and G_ℓ is k-degree-anonymous.

Note that a solution to the MIN ANONYMOUS-EDGE-ROTATION problem may not exist for all instances. For example, if G is a complete graph without an edge, $K_n \setminus \{e\}$, $n \geq 6$, then there is no solution for such a graph G and $k = 3$. Therefore, we are only interested in studying **feasible instances** (G, k) defined as instances for which there exists a solution to MIN ANONYMOUS-EDGE-ROTATION. Our initial study of sufficient conditions for feasibility is presented in Section 3.5.

Obviously, since all graphs are 1-degree-anonymous, we are only interested in cases where $k \geq 2$.

The decision version associated to MIN ANONYMOUS-EDGE-ROTATION is defined as follows for a feasible instance (G, k):

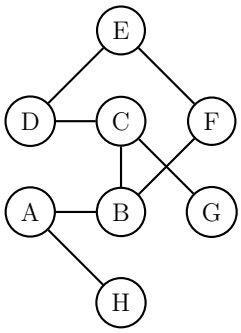
ANONYMOUS-EDGE-ROTATION
Input: (G, k, r) where $G = (V, E)$ is an undirected graph, $k \in \{1, \ldots, |V|\}$, and r be a positive integer.
Question: Is there a sequence of $\ell + 1$ graphs $G_0 = G, G_1, G_2, \ldots, G_\ell$ such that $\ell \leq r$, G_{i+1} can be obtained from G_i by one edge rotation, and G_ℓ is k-degree-anonymous?

We also consider the MIN ANONYMOUS-EDGE-ROTATION problem in restricted graph classes, e.g. trees. In that case we require that all graphs in the sequence $G_0, \ldots, G_\ell$ must be from the same graph class. Note that the problem can also be studied without this requirement, but the results may be different.

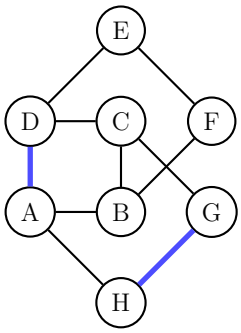
We will now summarize the different results on k-degree-anonymization. The notion was first studied in the literature through more classical operators like edge addition or deletion. Then we will focus on the work done with edge rotation.

Initially Liu and Terzi proposed to add or remove edges to reach a k-degree-anonymous graph [44]. Since a complete graph or an independent set is k-degree-anonymous for any $k \leq n$, we can transform any graph to a k-degree-anonymous one ($k \leq n$) starting for any graph we can always transform it into a k-degree-anonymous one by adding or deleting the minimum number of edges. Figure 2.2 gives an example of this process.

Subsequently, many papers have been published on this model. The table Table 2.1 summarizes the main results:



Vertex	Degree
A	2
B	3
C	3
D	2
E	2
F	2
G	1
H	1



Vertex	Degree
A	3
B	3
C	3
D	3
E	2
F	2
G	2
H	2

Figure 2.2: 4-degree-anonymization of a graph by edge additions (in blue)

Graph modif.	General Case		
	Exact	Approx.	Param. solution size
Edge addition	NP-hard [35]	Unknown	W[1]-hard [35]
Edge deletion	NP-hard [4]	$n^{1-\epsilon}$-inapprox. [4]	W[2]-hard [4]
Vertex deletion	NP-hard [4]	$n^{1-\epsilon}$-inapprox. [4]	W[2]-hard [4]
Vertex addition	NP-hard [12]	Unknown	FPT [12]

Table 2.1: The state of the art on vertex/edge addition/deletion

Vertex deletion In [4], Bazgan et al. proved that the problem is NP-hard, even on trees when $k = 2$. They also extended the reduction to trivially perfect graphs, bipartite permutation graphs and split graphs, still when $k = 2$. Still with the same reduction, they showed that the problem is not $n^{1-\epsilon}$-approximable in polynomial time (under $P \neq NP$), even on graphs with maximum degree three. The result holds even on FPT time parameterized by the solution size, $|S|$. On the other hand they showed that the problem is polynomial-time solvable on graphs of maximum degree two and on bounded degree cluster graphs. The problem has therefore been studied in depth and does not seem to be accessible for good algorithms with guaranteed performance.

Edge deletion In [4] the authors used a reduction similar to the one for vertex deletion and they deduced that the problem is NP-hard and there is no $n^{1-\epsilon}$-approximation, even in FPT time with respect to the size of the solution. There are still some tractable cases, for example when the solution size $|S|$ and k are given as parameters, the problem becomes FPT. As for vertex deletion, existing results show that there is little hope for an efficient guaranteed algorithm.

Edge addition There is a strong relation between edge addition and edge deletion: Finding the minimum number of edge deletions in order to make G k-degree-anonymous is equivalent to finding the minimum number of edge additions to make $\overline{G}$ k-degree-anonymous.
We can therefore conclude that all the results for edge deletion propagate to edge addition, on the complement of the concerned graph classes. Moreover, in [35], Hartung et al. studied this problem specifically. They showed that the problem is

NP-hard even on 3-colorable graphs with H-index three (i.e. at least 3 vertices of degree 3). Then they showed that the problem is $W[1]$-hard even if $k = 2$ and some FPT-time algorithms parameterized by the degree or the size of the solution.

Vertex addition For vertex addition the problem is a bit more complex, since it is necessary to define how the neighboring edges of the new vertex are added. In [12] the authors proposed different versions. They proved that the three versions are NP-hard. There are also many results on the parameterized complexity of the problem. Overall, two versions seem as hard as edge edits but the version where one chose the edges added look more accessible, it admits an FPT or XP algorithm for most parameters.

Edge move, edge rotation and edge switch The first of these graph modifications introduced under anonymity context is edge move. Hay et al. [36] proposed an anonymization algorithm based on a random edge move sequence and Ying et al. propose an improvement in [60]. Casas Roma et al. [15, 13] introduced edge move, edge rotation and edge switch in the context of k-degree-anonymization. They proposed heuristics to solve the problem as well as data utility comparisons between the different modifications. In a more theoretical context, Salas et al. [52] presented polynomial time algorithms in the case of edge rotations, however they minimize another metric than the number of rotations. There are no results about hardness.

Synthesis The theoretical works done on classical graph modifications (edge/vertex addition/deletion) showed that the problems are very difficult. On the other hand, the anonymization community was interested in different graph modifications. Therefore, it seemed appropriate to study these new modifications (edge move, edge rotation and edge switch) in order to better understand their complexity and perhaps obtain more encouraging results.

The chapter is organized as follows. The study of feasibility is initiated in Section 3.5. Section 3.5 presents the NP-hardness proof of Anonymous-Edge-Rotation. In Section 3.5 we study properties of the specific k-degree anonymous degree sequences that are used in Section 3.5 to present a polynomial-time 2-approximation algorithm and in Section 3.5 to establish a polynomial time algorithm for trees. Moreover in Section 3.5 we consider the case $k = n$ in general graphs. Some conclusions are given at the end of the chapter.

2.2 Feasibility

As it was discussed in Section 2.1, the Min Anonymous-Edge-Rotation problem does not have a solution for every input instance. It is not difficult to see that if a graph is 'almost' complete or 'almost' empty, then there are only restricted options on the number of different degree classes and therefore a solution may not exist.

First we show that we can reach any graph from any other graph with the same number of vertices and edges via edge rotations. Then we present some sufficient conditions for an instance to be feasible.

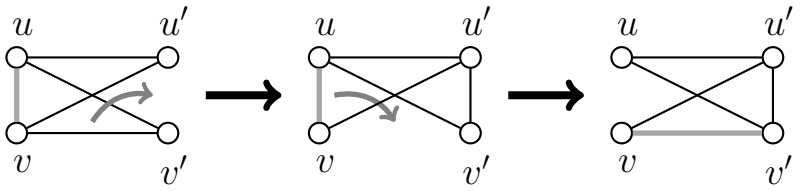


Figure 2.3: Case 1 of Theorem 3.5.1

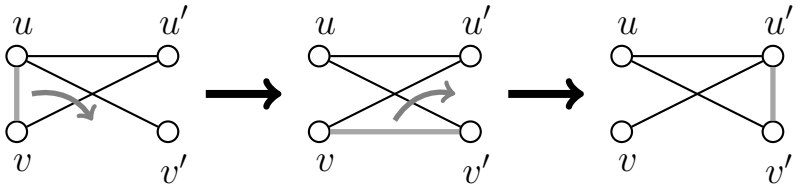


Figure 2.4: Case 2 of Theorem 3.5.1

The following theorem shows important properties about the edge rotations. The result was already proved in [16], but due to the simplicity of our approach, we present another proof here.

Theorem 2.2.1. *For any two graphs $G, G' \in \mathcal{G}(n, m)$, we can transform G into G' using a sequence of edge rotations.*

Proof. Let $E_1 = E(G) \setminus (E(G) \cap E(G'))$ be the set of edges that are in G and not in G' and $E_2 = E(G') \setminus (E(G) \cap E(G'))$ the set of edges that are in G' and not in G. For all u, v and w such as $uv \in E_1$ and $uw \in E_2$, we add one edge rotation (uv, uw). In all other cases, let $uv \in E_1$ and $u'v' \in E_2$, where all vertices u, v, u', v' are distinct. There are two cases: 1) uu', uv', vu' and $vv' \in E(G)$ or 2) at least one of these four edges is missing.

In the first case we can make the following two edge rotations to move uv from G to $u'v'$ in G': $(v'v, v'u')$ and (vu, vv') (see Figure 2.3). In the second case, if for example vv' is missing, we can use the following two rotations (vu, vv') and then $(v'v, v'u')$ (see Figure 2.4) and similarly if another edge is missing.

□

Corollary 2.2.1. *For any two graphs $G, G' \in \mathcal{G}(n, m)$, the edge rotation distance between G and G' is bounded by $2m$.*

Theorem 2.2.2. *Let $G \in \mathcal{G}(n, m)$ such that $\frac{n}{2} \leq m \leq \frac{n(n-3)}{2}$ and $n \geq 8$. Then there exists a feasible solution for the* Min Anonymous-Edge-Rotation *problem, hence a k-degree-anonymous graph $G' \in \mathcal{G}(n, m)$, for any $k \leq \frac{n}{4}$.*

Proof. Let m, n, k be fixed. Any graph $G \in \mathcal{G}(n, m)$ is a 1-degree-anonymous graph, hence we can suppose $k \geq 2$ and $n \geq 8$.

In the first part of the proof we describe a construction of a k-anonymous sequence $D = (d_1, d_2, \ldots, d_n)$ with property $\sum_{i=1}^{n} d_i = 2m$ for any m, n, k satisfying the restriction of the theorem. In the second part we show that the sequence D

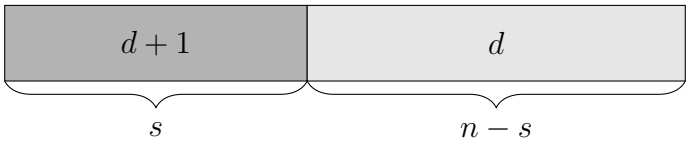


Figure 2.5: The sequence D_1

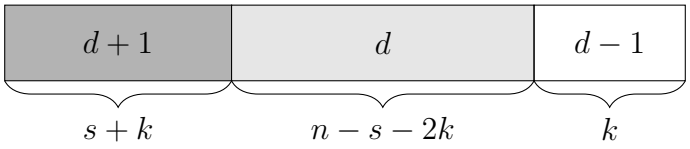


Figure 2.6: The sequence D_2

is graphic, hence that the sequence satisfies the conditions (1.1) and (1.2) from Section 1.1.

As $\sum_{i=1}^{n} d_i = 2m$ is the condition for a constructed sequence, the condition (1.1) trivially holds.

Now we construct three distinct k-anonymous sequences Type 1, 2, 3 of integers based on the values of k and $s \equiv 2m \bmod n$. Denote by d the average degree of the graph G defined as $d = \left\lfloor \frac{2m}{n} \right\rfloor$.

Type 1: $k \leq s \leq n-k$

Let $D_1 = (d_1^1, d_2^1, \ldots, d_s^1, d_1^2, d_2^2, \ldots, d_{n-s}^2)$ be a sequence of positive integers where for all i, $1 \leq i \leq s$, $d_i^1 = d+1$ and for all j, $1 \leq j \leq n-s$, $d_j^2 = d$ (see Figure 2.5). The sequence contains n elements and it is easy to see that $\sum_{i=1}^{s}(d+1) + \sum_{j=1}^{n-s} d = 2m$.
Following the assumptions $s \geq k$ and $n-s \geq k$, therefore D_1 is a k-anonymous sequence.

Type 2 : $s < k$

Let $D_2 = (d_1^1, d_2^1, \ldots, d_{s+k}^1, d_1^2, d_2^2, \ldots, d_{n-s-2k}^2, d_1^3, d_2^3, \ldots, d_k^3)$ be a sequence of positive integers where for all i, $1 \leq i \leq s+k$, $d_i^1 = d+1$; for all r, $1 \leq r \leq n-s-2k$, $d_r^2 = d$; for all j, $1 \leq j \leq k$, $d_j^3 = d-1$ (see Figure 2.6). The sequence contains n elements and $\sum_{i=1}^{s+k}(d+1) + \sum_{j=1}^{k}(d-1) + \sum_{\ell=1}^{n-s-2k} d = 2m$.
Since $n \geq 4k$ and $s < k$, $n-s-2k \geq k$, and thus D_2 is a k-anonymous sequence.

Type 3: $s > n-k$

Let $D_3 = (d_1^1, d_2^1, \ldots, d_k^1, d_1^2, d_2^2, \ldots, d_{s-2k}^2, d_1^3, d_2^3, \ldots, d_{k+n-s}^3)$ be a sequence of positive integers where for all i, $1 \leq i \leq k$, $d_i^1 = d+2$; for all r, $1 \leq r \leq s-2k$, $d_r^2 = d+1$; for all j, $1 \leq j \leq k+n-s$, $d_j^3 = d$ (see Figure 2.7). The sequence has n elements and $\sum_{i=1}^{k}(d+2) + \sum_{j=1}^{k+n-s} d + \sum_{\ell=1}^{s-2k}(d+1) = 2m$.
Because $n > s$, the number d appears more than k times in D_3. Due to the assumptions $n \geq 4k$ and $s > n-k$, we also have $s-2k \geq k$. Hence D_3 is a k-anonymous sequence.

Now we show that all three sequences are graphic, therefore that the condition

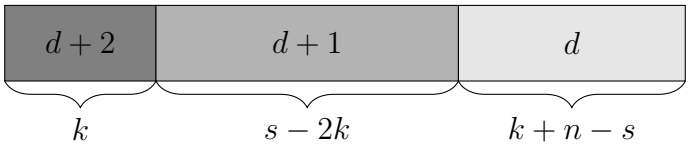


Figure 2.7: The sequence D_3

(1.2) is true for any ℓ. We split the proof into several sub-cases depending on the value of ℓ and the type of the sequence.

From our assumptions $\frac{n}{2} \leq m \leq \frac{n(n-3)}{2}$, it follows $1 \leq d \leq n-3$.

Case A: $\ell = 1$
Because $d \geq 1$, (1.2) trivially holds.

Case B: $\ell = 2$.
Since $n \geq 8$, $d \geq 1$ and $|\{x \in D_i \mid x = d-1\}| \leq |\{x \in D_i \mid x = d\}|$ for every $i \in \{1,2,3\}$ then $\sum_{j=\ell+1}^{n} \min(d_j, \ell) \geq 8d - 4 \geq 6d - 2$. We deduce:
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) = 2(d+2) \leq 2 + (6d-2) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Case C: $3 \leq \ell < d$
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) \leq \ell(n-1) = n\ell - \ell = \ell^2 - \ell + n\ell - \ell^2 = \ell(\ell-1) + (n-\ell)\ell \leq$
$\ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Case D: $3 \leq \ell = d$,
Type 1 & 3:
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) \leq \ell(n-1) = n\ell - \ell = \ell^2 - \ell + n\ell - \ell^2 = \ell(\ell-1) + (n-\ell)\ell \leq$
$\ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 2, following our assumptions we also have $\ell = d \leq n-3$
$\sum_{i=1}^{\ell} d_i \leq \ell(d+1) = \ell(\ell+1) = \ell(\ell-1) + 2\ell \leq \ell(\ell-1) + 3\ell - 3 = \ell(\ell-1) + 3(\ell-1) =$
$\ell(\ell-1) + (\ell+3-\ell)(\ell-1) \leq \ell(\ell-1) + (n-\ell)(\ell-1) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Case E: $3 \leq \ell = d+1$. Furthermore, $\ell = d+1 \leq n-2$.

Type 1 & 2, $\ell \geq 4$:
$\sum_{i=1}^{\ell} d_i \leq \ell(d+1) = \ell^2 = \ell(\ell-1) + \ell \leq \ell(\ell-1) + 2\ell - 4 = \ell(\ell-1) + 2(\ell-2) =$
$\ell(\ell-1) + (\ell+2-\ell)(\ell-2) \leq \ell(\ell-1) + (n-\ell)(\ell-2) = \ell(\ell-1) + (n-\ell)(d-1) \leq$
$\ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 1 & 2, $\ell = 3$:
Due to $n \geq 8$, $\sum_{j=\ell+1}^{n} \min(d_j, \ell) \geq 5d - 4 \geq \ell$
Therefore $\sum_{i=1}^{\ell} d_i \leq \ell(d+1) = \ell^2 = \ell(\ell-1) + \ell \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 3, $3 \leq \ell \leq n-3$
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) = \ell(\ell+1) = \ell(\ell-1) + 2\ell \leq \ell(\ell-1) + 3\ell - 3 = \ell(\ell-1) + 3(\ell-1) =$
$\ell(\ell-1) + (\ell+3-\ell)(\ell-1) \leq \ell(\ell-1) + (n-\ell)(\ell-1) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 3, $\ell = n-2$
$\sum_{i=1}^{\ell} d_i = k(d+2) + (s-2k)(d+1) + (k+n-s-2)d = nd - 2d + s \leq d(n-2) + n - 1 =$
$(n-3)(n-2) + n - 1 \leq \ell(\ell-1) + 2d = \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell).$

Case F: $3 \leq \ell = d+2$. Furthermore, $\ell = d+2 \leq n-1$.
Type 1 & 2:
$\sum_{i=1}^{\ell} d_i \leq \ell(d+1) = \ell(\ell-1) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 3, $\ell = 3$:
Due to $n \geq 8$, $\sum_{j=\ell+1}^{n} \min(d_j, \ell) \geq 5 \geq \ell$. Then
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) = \ell^2 = \ell(\ell-1) + \ell \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 3, $4 \leq \ell \leq n-2$:
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) = \ell^2 = \ell(\ell-1) + \ell \leq \ell(\ell-1) + 2\ell - 4 = \ell(\ell-1) + 2(\ell-2) =$
$\ell(\ell-1) + (\ell+2-\ell)(\ell-2) \leq \ell(\ell-1) + (n-\ell)(\ell-2) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Type 3, $\ell = n-1$:
$\sum_{i=1}^{\ell} d_i = k(d+2) + (s-2k)(d+1) + (k+n-s-1)d = s + nd - d \leq n - 1 + (n-1)(n-3) =$
$(n-1)(n-2) = \ell(\ell-1) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell).$

Case G: $d+2 < \ell < n$
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) \leq \ell(\ell-1) \leq \ell(\ell-1) + \sum_{j=\ell+1}^{n} \min(d_j, \ell)$

Case H: $\ell = n$
$\sum_{i=1}^{\ell} d_i \leq \ell(d+2) \leq \ell(\ell-1)$

Therefore, we have proved that there exists a k-degree-anonymous graph $G' \in \mathcal{G}(n, m)$ and the graph G can be transformed into G' using a sequence of edge rotations thanks to Theorem 3.5.1. □

Now we extend the feasibility study to the case $k = n$ for which we get necessary and sufficient conditions.

Theorem 2.2.3. *Let $G \in \mathcal{G}(n, m)$ for some positive integers n and m. Then (G, n) is a feasible instance of* Min Anonymous-Edge-Rotation *if and only if $\frac{2m}{n}$ is an integer.*

Proof. Since $k = n$ in Min Anonymous-Edge-Rotation, every vertex has to be in the same degree class, so if there is a solution, the resulting graph has to be regular. Moreover, a necessary and sufficient condition for a p-regular graph with n vertices to exist is that $n \geq p + 1$ and np must be even [57].

If $\frac{2m}{n}$ is not an integer then obviously there is no regular graph in $\mathcal{G}(n, m)$ and therefore (G, n) is not a feasible instance.

If $\frac{2m}{n}$ is an integer, since $n \times \frac{2m}{n} = 2m$ is even and $n \geq \frac{2m}{n} + 1$ there is a $\frac{2m}{n}$-regular graph in $\mathcal{G}(n, m)$ as it was mentioned before. By Theorem 3.5.1 we conclude that there exists a sequence of edge rotations that leads to a $\frac{2m}{n}$-regular graph starting from G. □

2.3 NP-hardness

In this section we show that the decision version of Min Anonymous-Edge-Rotation, the problem Anonymous-Edge-Rotation, is NP-hard. The proof is based on a reduction from the restricted version of a cover set problem, Restricted Exact Cover By 3-Sets, which is known to be NP-complete [27].

Restricted Exact Cover By 3-Sets (RX3C)
Input: A set X of elements with $\|X\| = 3m$ and a collection C of 3-elements subsets of X where each element appears in exactly 3 sets.
Question: Does C contain an exact cover for X, i.e. a subcollection $C' \subseteq C$ such that every element occurs in exactly one member set of C' ?

Remark 2.3.1. *Note that $\|C\| = 3m$ and we can suppose that m is even and larger than 6. If m is odd, we consider the instance I_{even} defined as follows: $X_{even} = X \cup \{x' \mid x \in X\}$ and $C_{even} = C \cup \{c_{x'y'z'} \mid c_{xyz} \in C\}$, and thus in the new instance I_{even} the set has $6m$ elements and the collection has $6m$ 3-elements subsets.*

We define a polynomial-time reduction and then prove the NP-hardness of Anonymous-Edge-Rotation.

Reduction. Let $I = (X, C)$ be an instance of $RX3C$ with $|X| = |C| = 3m$ and m even and $q \geq 3$ a given constant. We describe the construction σ transforming an instance I into the graph $G := \sigma(I)$ where $G = (V, E)$ is defined as follows:

- For each element $x \in X$, we add a vertex v_x to the set $V_{elem} \subset V$ and a vertex u_x to the set $V_{hub} \subset V$.
- For each 3-element set $\{x, y, z\}$ of the collection C, we add 4 vertices c^1_{xyz}, c^2_{xyz}, c^3_{xyz} and c^4_{xyz} to the set $V_{set} \subset V$.

- For each $i \in \{1, \ldots, 5m\}$ we add a vertex w_i to the set $V_{reg} \subset V$ and for each $j \in \{1, \ldots, 10m\}$ we add a vertex t_j to $V_{single} \subset V$.

Let $V^- = V_{elem} \cup V_{hub} \cup V_{set} \cup V_{reg} \cup V_{single}$ and $|V^-| = 3m+3m+12m+15m = 33m$. If $q = 3$, then let $V = V^-$. If $q \geq 4$, then for each i, $4 \leq i \leq q$, add a set of $11m$ vertices denoted V^i_{dummy}. Let $V_{dummy} = V^4_{dummy} \cup \cdots \cup V^q_{dummy}$ and define $V = V^- \cup V_{dummy}$. Obviously, $|V| = 33m + (q-3)11m$.

Now we define the set E of the edges in G.

- For all $x, y \in X$, such that $x \neq y$, we add the edge $v_x u_y$ between the vertex $v_x \in V_{elem}$ and $u_y \in V_{hub}$, to $E_X \subset E$.
- For each 3-element set $\{x, y, z\}$ of the collection C, $\forall i \in \{1, 2, 3, 4\}$, we add the edges $c^i_{xyz}u_x$, $c^i_{xyz}u_y$ and $c^i_{xyz}u_z$ to the set $E_C \subset E$.
- We add the set of edges $E' \subset E$ to the vertex set V_{elem} such that (V_{elem}, E') is a 11-regular graph. Since the number of vertices in the set $|V_{elem}| = 3m$ is even (m is even) and $11 < 3m$ such a regular graph exists [57]. Furthermore, such a graph can be constructed in polynomial time using Havel-Hakimi algorithm [33].
- We add the set of the edges $E'' \subset E$ to the vertex set V_{reg} such that (V_{reg}, E'') is a $(3m + 11)$-regular graph. Since the number of vertices of V_{reg} is even and $3m + 11 < 5m$, similarly to the previous case such a regular graph exists and can be constructed in polynomial time.

Finally, let $E^- = E_X \cup E_C \cup E' \cup E''$. If $q = 3$, then let $E = E^-$. If $q \geq 4$, then the set E contains E^- and for any i, such that $4 \leq i \leq q$, we add the set of edges $E^i_{dummy} \subseteq E$ to the vertex set V^i_{dummy} such that $(V^i_{dummy}, E^i_{dummy})$ is $(9m + 12)$-regular. Since the number of vertices of V^i_{dummy} is even (m is even) and $9m+12 \leq 11m$, similarly to the previous case such a regular graph exists and can be constructed in polynomial time.

Obviously, the graph $G = (V, E)$ has the following properties: (i) $10m$ vertices of degree 0 (the vertices of the set V_{single}), (ii) $12m$ vertices of degree 3 (the vertices of the set V_{set}), (iii) $8m$ vertices of degree $3m + 11$ (the vertices of the set V_{reg} and V_{hub}), (iv) $3m$ vertices of degree $3m+10$ (the vertices of the set V_{elem}), (v) $(q-3)11m$ vertices of degree $(9m + 12)$ (the vertices of the set V_{dummy}).

Remark 2.3.2. *Let us note that $n = q \times 11m$.*

Example. Figure 2.8 represents the transformation σ for $q = 3$. Let I_1 be the following instance of X3C: $m = 2$, $X = \{1, 2, 3, 4, 5, 6\}$, and $C = \{\{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}, \{4, 5, 6\}, \{1, 5, 6\}, \{1, 2, 6\}\}$. To simplify the figure, we only consider $m = 2$, but for the construction m must be at least 6 (due to an $(3m + 11)$-regular graph on the vertex set of V_{reg}).

Theorem 2.3.1. ANONYMOUS-EDGE-ROTATION *is NP-complete even in case $k = \frac{n}{q}$ where n is the order of the graph G for an input instance (G, k, r) and q is a fixed number greater than or equal to 3.*

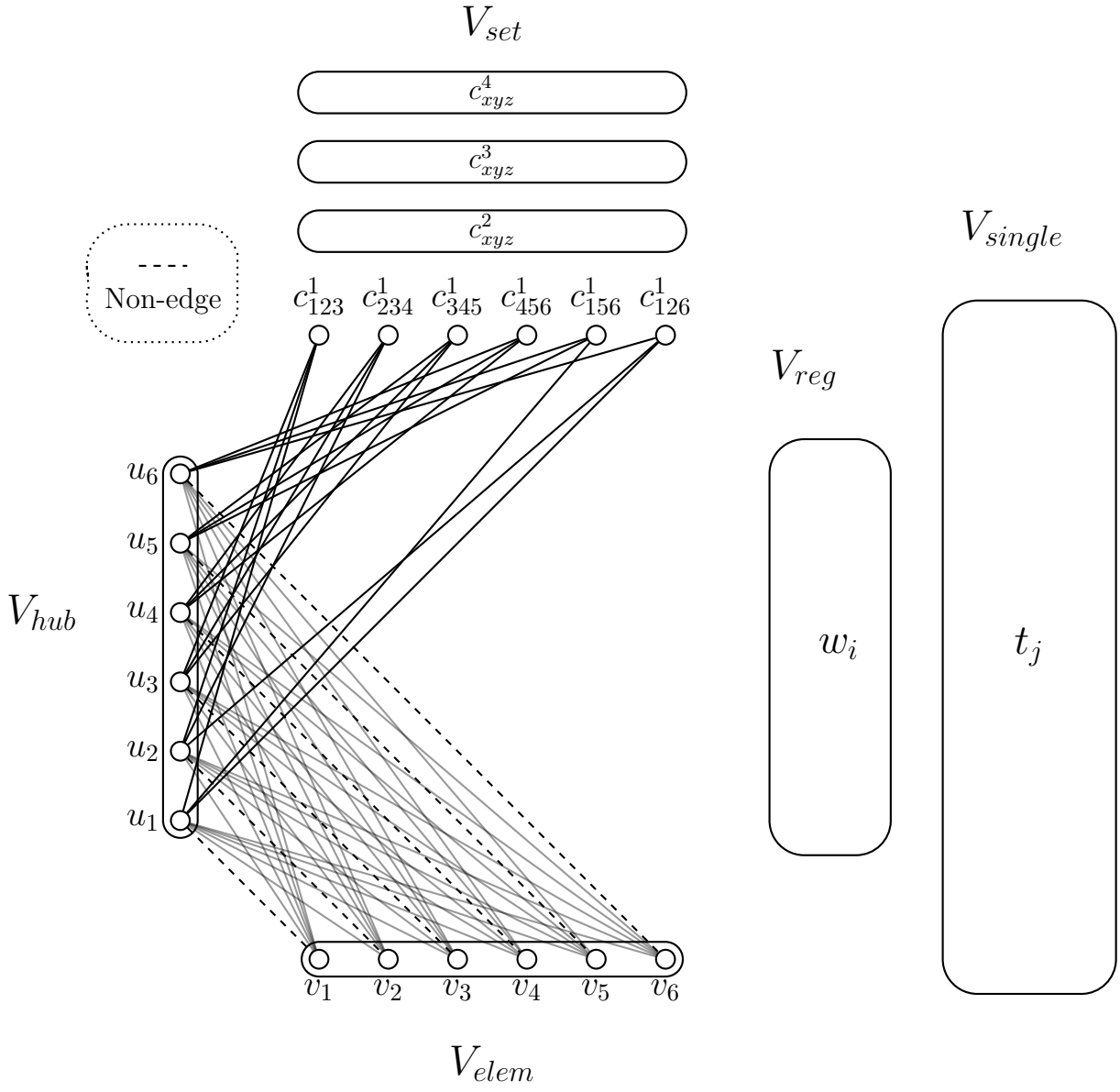


Figure 2.8: $\sigma(I_1)$

Proof. Let $C' \subseteq C$ be an exact cover for X of size m. Now we define $r = 3m$ rotations which are independent from each other : for every 3-element set $\{x, y, z\} \in C'$, we replace the edge $u_x c^1_{xyz}$ by the edge $u_x v_x$, and similarly $u_y c^1_{xyz}$ by $u_y v_y$ and $u_z c^1_{xyz}$ by $u_z v_z$. Since C' is of size m, we define exactly $3m$ rotations. Let G' be the graph obtained from G after applying all $3m$ rotations. Since C' is an exact cover of size m: (i) there are m vertices of type c^1_{xyz} that lost all 3 neighbours and become of degree 0 in G', (ii) all $3m$ vertices of type v_x are attached to a new neighbour, so they become of degree $3m + 11$ in G'.

Then G' has $10m + m = 11m$ vertices of degree 0, $12m - m = 11m$ of degree 3 vertices, $8m+3m = 11m$ of degree $3m+11$ vertices and it contains $q-3$ disconnected $(9m + 12)$-regular subgraphs of size $11m$, hence we conclude that G' is a $11m$-anonymous graph.

Let I' be a yes-instance of Anonymous-Edge-Rotation. Then there exists a sequence of $r = 3m$ rotations such that the graph $G' = (V, E')$ obtained after applying the rotations to G is a $11m$-anonymous graph. Since $|V| = 33m + (q - 3)11m$, there must be only q different degrees classes in G'. Note that with one rotation, we can change the degree of two vertices, therefore the degree at most $6m$ vertices can be changed by $3m$ rotations. Since the graph G has more than $6m$ vertices of the degrees $3m + 11$, 3, 0 and $9m + 12$, all these degree classes must be in G'. Furthermore, due to the number of vertices of G, these are the only degree classes in G'. This means that in G' the number of vertices of degree $3m + 11$ must be increased by $3m$, the number of vertices of degree 0 must be increased by m, the number of vertices of degree 3 must be decreased by m, there are no vertices of degrees $3m+10$ in G' and the other degree classes keep the same amount of vertices.

A single rotation can increase or decrease the degree of a vertex by 1 therefore using $3m$ rotations no vertex of degree $3m + 10$ in G can have degree 0 in G' and similarly, no vertex of degree 3 in G can have degree $3m + 11$ in G'. Therefore the $3m$ new vertices of degree $3m+11$ in G' must have degree $3m+10$ in G. This is only possible if the degree of each vertex v_x from the set V_{elem} is increased by 1. Similarly, the m new vertices of degree 0 in G' must have degree 3 in G, let $C_{G'}$ be the set of such vertices. Obviously, $C_{G'}$ must be a subset of V_{set}, in which the vertices have the form c^ℓ_{xyz} with $x, y, z \in X$, for any set $\{x, y, z\} \in C$, and $\ell \in \{1, 2, 3, 4\}$. For the same reasons, vertices of degree greater than $9m + 12$ in G cannot be of degree less than $3m + 12$ in G'.

To reach the requested degree configuration in G' with exactly $3m$ edge rotations, in each rotation the degree of each vertex from V_{elem} must be increased by 1 and the degree of each vertex from the set $C_{G'}$ must be decreased by 1. To achieve that, for each vertex v_x from V_{elem}, the only possible rotation is to add the edge $u_x v_x$ where $u_x \in V_{hub}$ and remove the edge $u_x c^\ell_{xyz}$ where $c_{xyz} \in C_{G'}$. To fulfil the condition about the degree classes and the number of the rotations, the only way to achieve that is that $C'' = \{\{x, y, z\} \mid c^\ell_{xyz} \in C_{G'}\}$ is an exact cover of X. □

2.4 Lower bound for rotations

In this section we suppose that (G, k) is a feasible instance. For any such instance we define a k-anonymous degree sequence S_{bound} that can be computed in polynomial time if $k = \Theta(n)$. We show that with the $(+1, -1)$-degree modifications (Remark 3.5.1) the graph G can be transformed into a k-degree-anonymous graph G' with degree sequence S_{bound} using at most double of edge rotations as in an optimal solution of Min Anonymous-Edge-Rotation for (G, k).

Note that in general a $(+1, -1)$-degree modification does not correspond to an edge rotation, but as we show later in Section 3.5, it is true for trees.

Now in the following steps we show how to define the degree sequence S_{bound}.

Step 1: Compute every available target sequence

Let $S = (s_1, \ldots, s_n)$ be a non-increasing sequence of non-negative integers, $r \in \{1, \ldots, n\}$. Any partition of S into r contiguous subsequences (i.e. if $S[a]$ and $S[b]$ are in one part, then all $S[i]$, $a \leq i \leq b$ must be in the same part) is called a contiguous r-partition. The number of contiguous r-partitions of S is $\binom{n-1}{r-1}$, therefore bounded by $(n-1)^{r-1}$. Then the number of contiguous partitions of S with at most r parts can be bounded by $\sum_{i=0}^{r-1}(n-1)^i \leq 2n^{r-1}$.

For each contiguous ℓ-partition p, $1 \leq \ell \leq r$, we use notation $p = [p_1, \ldots, p_\ell]$, where p_i denotes the number of elements in part i, $1 \leq i \leq \ell$. Note that at this stage what is important is the number of elements in each part, not which elements from S are in it.

Let G be a graph of order n and k an integer, $k \geq 2$. If G is a k-degree-anonymous graph, then the vertices of G can be partitioned into at most $c = \lfloor \frac{n}{k} \rfloor$ parts where the vertices in each part have the same degree. Let P be the set of all such contiguous partitions with at most c parts. As it follows from the initial discussion, the number of such partitions is bounded by $2n^{c-1}$.

Now for each contiguous partition $p = [p_1, p_2, \ldots, p_\ell] \in P$, $\ell \in \{1, \ldots, c\}$, we compute all non-increasing sequences $(d_1, d_2, \ldots, d_\ell)$ of ℓ integers d_i such that $0 \leq d_i < |V|$. Let $\hat{P}_p$ be the set of all feasible k-anonymous degree sequences for p, i.e.

$$S = (\underbrace{d_1, \ldots, d_1}_{p_1-\text{times}}, \underbrace{d_2, \ldots, d_2}_{p_2-\text{times}}, \ldots, \underbrace{d_\ell, \ldots, d_\ell}_{p_\ell-\text{times}}) = (d_1^{p_1}, d_2^{p_2}, \ldots, d_\ell^{p_\ell}) \in \hat{P}_p$$

if and only if $\sum_{i=1}^{\ell} p_i d_i = 2|E|$, S is graphic and k-anonymous.

For each contiguous partition p with ℓ parts, $1 \leq \ell \leq c$, there are at most n possibilities for a degree in each position. The test whether the generated sequence is graphic and k-anonymous can be done in $O(n)$ operations. Since $|P| = O(n^{c-1})$, there are at most $O(n^{c-1} \times n^\ell \times n) \leq O(n^{2c})$ operations to compute all feasible degree sequences of every partition, where $c = \lfloor \frac{n}{k} \rfloor$. Obviously, if c is a constant, such a number of operations is polynomial.

Step 2: Find the best one

Now based on the previous analysis we can define the degree sequence S_{bound} and prove some basic properties.

Definition 2.4.1. *Let G be a graph of order n with degree sequence S_G. Then define S_{bound} for G as a degree sequence for which the sum $\sum_{i=1}^{n} |S_G[i] - S[i]|$ achieves the minimum over all elements $S \in \hat{P}_p$ and $p \in P$.*

Remark 2.4.1. Similarly to a k-anonymous sequence S_{bound} defined in Definition 2.4.1 for a graph, we can define a k-anonymous sequence S_{Tbound} for a tree. The only difference is that in the set $\hat{P}_p$, every feasible solution must have $d_i \geq 1$, which would be a subset of $\hat{P}_p$. Also for the testing, we do not need to check whether S is graphic, the condition $\sum_{i=1}^{\ell} p_i d_i = 2|E|$, is enough for the degree sequence of a tree.

Lemma 2.4.1. *Let S be a n-sequence of non-negative integers and denote by S' the sequence S sorted in non-increasing order. Let S_s be another n-sequence of non-negative integers sorted in non-increasing order. Then*

$$\sum_{i=1}^{n} |S_s[i] - S'[i]| \leq \sum_{i=1}^{n} |S_s[i] - S[i]| \tag{2.1}$$

Proof. If S is already in non-increasing order then (2.1) holds. If not then there exist positive integers a, b such that $a < b$ and $S[a] < S[b]$. Let S_1 be the sequence defined swapping the values $S[a]$, $S[b]$, hence: $S_1[a] = S[b]$, $S_1[b] = S[a]$, and $S_1[i] = S[i]$ otherwise. We denote

$$\begin{aligned} A &= \sum_{i=1}^{n} |S_s[i] - S[i]| - \sum_{i=1}^{n} |S_s[i] - S_1[i]| \\ &= |S_s[a] - S[a]| - |S_s[a] - S_1[a]| + |S_s[b] - S[b]| - |S_s[b] - S_1[b]| \\ &= |S_s[a] - S_1[b]| - |S_s[a] - S_1[a]| + |S_s[b] - S_1[a]| - |S_s[b] - S_1[b]| \end{aligned}$$

In order to follow easier six different cases, let $x_1 = S_s[a]$, $x_2 = S_s[b]$, $x_3 = S_1[a]$, $x_4 = S_1[b]$, and thus $A = |x_1 - x_4| - |x_1 - x_3| + |x_2 - x_3| - |x_2 - x_4|$. Following our assumptions $x_1 \geq x_2$ and $x_3 > x_4$.

Now for all possible arrangements of x_1, x_2, x_3, x_4 we discuss the value A:

- $x_1 \geq x_2 \geq x_3 > x_4 : A = x_1 - x_4 - x_1 + x_3 + x_2 - x_3 - x_2 + x_4 = 0$
- $x_3 > x_4 \geq x_1 \geq x_2 : A = x_4 - x_1 - x_3 + x_1 + x_3 - x_2 - x_4 + x_2 = 0$
- $x_1 \geq x_3 > x_4 \geq x_2 : A = x_1 - x_4 - x_1 + x_3 + x_3 - x_2 - x_4 + x_2 = 2x_3 - 2x_4 > 0$
- $x_3 \geq x_1 \geq x_2 \geq x_4 : A = x_1 - x_4 - x_3 + x_1 + x_3 - x_2 - x_2 + x_4 = 2x_1 - 2x_2 \geq 0$
- $x_1 \geq x_3 \geq x_2 \geq x_4 : A = x_1 - x_4 - x_1 + x_3 + x_3 - x_2 - x_2 + x_4 = 2x_3 - 2x_2 \geq 0$
- $x_3 \geq x_1 \geq x_4 \geq x_2 : A = x_1 - x_4 - x_3 + x_1 + x_3 - x_2 - x_4 + x_2 = 2x_1 - 2x_4 \geq 0$

We can conclude that in all cases $A \geq 0$, therefore $\sum_{i=1}^{n} |S_s[i] - S_1[i]| \leq \sum_{i=1}^{n} |S_s[i] - S[i]|$.

If the sequence S_1 is still not in non-increasing order, we can repeat the process of swapping for the next two unsorted elements on S_1 until we obtain the non-increasing sequence S'. Each process can be repeated independently, therefore

$$\sum_{i=1}^{n} |S_s[i] - S'[i]| \leq \sum_{i=1}^{n} |S_s[i] - S[i]|.$$

□

Theorem 2.4.1. *Let (G, k) be a feasible instance for the* Min Anonymous-Edge-Rotation *problem. Let OPT be an optimum solution that is a minimum set of rotations that transform G into a k-degree-anonymous graph G'. Then $\sum_{i=1}^{n} |S_G[i] - S_{bound}[i]| \leq 2|OPT|$, where the degree sequence S_{bound} is defined in Definition 2.4.1.*

Proof. Let $S_{G'}$ be the degree sequence of G' sorted in the same order as S_G (i.e. for every $v \in V$, if $d_G(v)$ is in the position i in S_G then $d_{G'}(v)$ is in the position i in $S_{G'}$). Let $S'_{G'}$ be the degree sequence $S_{G'}$ sorted in non-increasing order. As in the definition of S_{bound} we considered all the options, there must exist $p \in P$ and $S \in \hat{P}_p$ such that $S = S'_{G'}$, and

$$\sum_{i=1}^{n} |S_G[i] - S_{bound}[i]| \leq \sum_{i=1}^{n} |S_G[i] - S'_{G'}[i]|.$$

Since the degree sequence $S'_{G'}$ is sorted in non-increasing order, then

$$\sum_{i=1}^{n} |S_G[i] - S'_{G'}[i]| \leq \sum_{i=1}^{n} |S_G[i] - S_{G'}[i]|$$

by Lemma 2.4.1. One rotation from the graph G_j to G_{j+1} in the sequence of the graphs from G to G' can only decrease the degree of a vertex by one and increase the degree of another one by one, hence $\sum_{i=1}^{n} |S_{G_j}[i] - S_{G'}[i]| \leq \sum_{i=1}^{n} |S_{G_{j+1}}[i] - S_{G'}[i]| + 2$. This means by one rotation the value $\sum_{i=1}^{n} |S_G[i] - S_{G'}[i]$ decreases by at most 2. After $|OPT|$ rotations, the last graph G_{j+1} in the sequence is G', therefore $\sum_{i=1}^{n} |S_G[i] - S_{G'}[i]| \leq 2|OPT|$ and the theorem follows. □

2.5 Approximation

In this section we show that under some constraints on the number of edges and k, there exists a polynomial time 2-approximation algorithm for the Min Anonymous-Edge-Rotation problem for all feasible inputs (G, k).

Remark 2.5.1. Let $S = (x_1, x_2, \ldots, x_n)$ be a non-increasing sequence of n non-negative integers. Denote by $R = x_1 - x_n$, $A_0 = \frac{x_1 + x_n}{2}$, and let $A = \frac{\sum_{i=1}^{n} x_i}{n}$.

The standard deviation of S is defined as $\sigma(S) = \sqrt{\frac{\sum(x_i-A)^2}{n}}$. It can be shown that

$$\sum_{i=1}^{n}(x_i - A)^2 \leq \sum_{i=1}^{n}(x_i - A_0)^2 \leq \frac{nR^2}{4},$$

hence $\sigma(S) \leq \frac{R}{2}$.

The mean absolute deviation of S is defined as $MAD[S] = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$. It is well known (e.g. applying Jensen's inequality) that $MAD[S] \leq \sigma(S)$.

Based on the correlation mentioned in Remark 2.5.1, we calculate an upper bound on the values in the degree sequence S_{bound} in the following lemma.

Lemma 2.5.1. *Let (G, k) be an instance of the* Min Anonymous-Edge-Rotation *problem where G is the graph with n vertices and m edges. Suppose that $\frac{n}{2} \leq m \leq \frac{n(n-3)}{2}$, $k \leq \frac{n}{4}$, and let the constant c be defined as $c = \lfloor \frac{n}{k} \rfloor$, hence $k = \Theta(n)$. Let S_{bound} be the k-anonymous degree sequence associated with G defined following Definition 2.4.1. Then for every i, $S_{bound}[i] \leq \min\{(1+\frac{n}{4k}+\frac{n}{k\Delta})\Delta, n-1\}$, $1 \leq i \leq n$.*

Proof. Let S_G be the degree sequence of G sorted in non-increasing order and D the k-anonymous degree sequence constructed following Theorem 3.5.2. Denote the unrounded average degree as $A = \frac{\sum_{i=1}^{n} S_G[i]}{n}$. Then using Remark 2.5.1, the standard deviation of S_G, $\sigma[S_G] \leq \frac{\Delta}{2}$, and $MAD[S_G] \leq \sigma(S_G)$. Hence

$$\begin{aligned}\sum_{i=1}^{n}|S_G[i] - D[i]| &\leq \sum_{i=1}^{n}\max(|S_G[i] - \lfloor A+2 \rfloor|, |S_G[i] - \lfloor A-1 \rfloor|) \\ &\leq \sum_{i=1}^{n}|S_G[i] - A| + \sum_{i=1}^{n}2 = nMAD[S_G] + 2n \\ &\leq n\sigma[S_G] + 2n \leq n\frac{\Delta}{2} + 2n = \frac{n(\Delta+4)}{2}\end{aligned}$$

Let Δ' be the maximum value of S_{bound}. If $\Delta' \leq \Delta$, then the condition from Lemma holds. If $\Delta' > \Delta$, then the distance between the k first elements of S_{bound} and the k first elements of S_G is at least $k(\Delta' - \Delta)$ since S_{bound} is k-anonymous and sorted in non-increasing order. Because $\sum_{i=1}^{n} S_{bound}[i] = \sum_{i=1}^{n} S_G[i]$, if the value of some elements is increased by a certain amount, the value of some others have to be decreased by the same amount, so $\sum_{i=1}^{n}|S_G[i] - S_{bound}[i]| \geq 2k(\Delta' - \Delta)$.

If $\Delta' > (1 + \frac{n}{4k} + \frac{n}{k\Delta})\Delta$ then $\sum_{i=1}^{n}|S_G[i] - S_{bound}[i]| > 2k(\frac{n}{4k} + \frac{n}{k\Delta})\Delta = \frac{n(\Delta+4)}{2} \geq \sum_{i=1}^{n}|S_G[i] - D[i]|$, which is not possible due to the minimality of S_{bound}. □

In the following two lemmas we prove that if a graph has 'sufficiently' many edges then edge rotations with the specific properties exist in a graph.

Lemma 2.5.2. *Let $G = (V, E)$ be a graph with $|E| > \Delta^2$ and let $uv \in E$. Then there exists an edge $ab \in E$ such that both vertices a and b are different from u and v and at most one of the following edges $\{av, au, bv, bu\}$ is in E.*

Proof. For an edge $xy \in E$, let $N_x = \mathcal{N}_G(x) \setminus \{y\}$ and $N_y = \mathcal{N}_G(y) \setminus \{x\}$. For a contradiction suppose there exists an edge $uv \in E$ such that for every edge $ab \in E \setminus (Inc(u) \cup Inc(v))$ at least two of the edges $\{av, au, bv, bu\}$ are in E. Then at least one vertex from $\{a, b\}$ is incident to both vertices u, v, hence belongs to $N_u \cap N_v$, or both vertices $\{a, b\}$ are in $(N_u \cup N_v) \setminus (N_u \cap N_v)$. Moreover, every vertex in $N_u \cup N_v$ has at most $\Delta - 1$ neighbours in $V \setminus \{u, v\}$. Hence,

$$\begin{aligned}
|E \setminus (Inc(u) \cup Inc(v))| &\leq (\Delta - 1) \times (|N_u \cap N_v| + \frac{|(N_u \cup N_v) \setminus (N_u \cap N_v)|}{2}) \\
&= (\Delta - 1) \times \frac{|N_u \cap N_v| + |N_u \cup N_v)|}{2} \\
&= (\Delta - 1) \times \frac{|N_u| + |N_v|}{2} \\
&\leq (\Delta - 1)^2
\end{aligned}$$

Then $|E| \leq |Inc(u) \cup Inc(v)| + |E \setminus (Inc(u) \cup Inc(v))| \leq 1 + 2(\Delta - 1) + (\Delta - 1)^2 = \Delta^2$. This is in contradiction with the hypothesis $|E| > \Delta^2$. □

Lemma 2.5.3. *Let $G = (V, E)$ be a graph and suppose $|E| > \Delta^2$. Let $v^+, v^- \in V$ such that $1 \leq d_G(v^-) \leq \Delta$ and $0 \leq d_G(v^+) \leq \Delta < |V| - 1$. Then there exists a sequence of at most two edge rotations that transform G into G' such that $d_{G'}(v^+) = d_G(v^+) + 1$, $d_{G'}(v^-) = d_G(v^-) - 1$ and degrees of other vertices in G are not changed. These rotations can be found in $O(|E|)$ steps.*

Proof. Case 1: Suppose there exists a vertex $v \in V$ such that $v \in \mathcal{N}_G(v^-)$ and $v \notin \mathcal{N}_G(v^+)$. Let G' be the graph obtained from G removing the edge v^-v and adding the edge vv^+, hence using rotation (vv^-, vv^+). Obviously, $d_{G'}(v^+) = d_G(v^+) + 1$, $d_{G'}(v^-) = d_G(v^-) - 1$ and G' is obtained by using a single rotation.
Case 2 : $\mathcal{N}(v^-) \subseteq \mathcal{N}(v^+)$. Let $u \in \mathcal{N}_G(v^-)$. Since $|E| > \Delta^2$ and $uv^+ \in E$, by using Lemma 2.5.2 there exists an edge $ab \neq uv^+ \in E$ such that at most one edge of the set $\{av^+, au, bv^+, bu\}$ is in E. If au is in E (or none of the 4), then the graph G' obtained by two rotations (ab, av^+) and (uv^-, ub) has the required properties. If av^+ is in E, then the graph G' obtained by two rotations (ba, bv^+) and (uv^-, ua) has the required properties. The remaining two cases if bu or bv^+ are in E are symmetrical to the above cases, as it is enough to swap a and b.

Obviously, such an edge ab can be found in $O(|E|)$. □

Theorem 2.5.1. *The* Min Anonymous-Edge-Rotation *problem is polynomial time 2-approximable for all instances (G, k), $k \leq \frac{n}{4}$ where $k = \Theta(n)$ and $G \in \mathcal{G}(n, m)$, where $\max\{\frac{n}{2}, (1 + \frac{n}{4k} + \frac{n}{k\Delta})^2\Delta^2\} \leq m \leq \frac{n(n-3)}{2}$.*

Proof. Let $(G = (V, E), k)$ be an instance of Min Anonymous-Edge-Rotation and S_G be the degree sequence of G. Let the constant c be defined as $c = \lfloor \frac{n}{k} \rfloor$. Due to our assumptions about the number of edges and k, all such instances are feasible if $n \geq 8$ as follows from Theorem 3.5.2. First we compute a k-anonymous degree sequence S_{bound} following Definition 2.4.1 in $O(n^{2c})$ steps. Due to the assumption $k = \Theta(n)$ and consequently c being a constant, such a number of steps is polynomial.

Furthermore, the condition on the number of edges ensures that we can always apply Lemma 2.5.3 and find suitable edge rotations.

If there exist two vertices v^+, $v^- \in V$ such that $0 \leq S_G[v^+] < S_{bound}[v^+]$ and $S_G[v^-] > S_{bound}[v^-]$ we apply Lemma 2.5.3 to transform G into a graph G_1 with at most two rotations such that $d_{G_1}(v^+) = d_G(v^+) + 1$ and $d_{G_1}(v^-) = d_G(v^-) - 1$. If there are no such vertices then $S_{bound}[u] = S_G[u]$ for all $u \in V$.

We will be executing the above transformations while there are two vertices v^+, $v^- \in V$ with the required properties. In each such transformation we decrease the degree of one vertex by 1 and increase the degree of another one by 1 with at most two rotations. Hence we transform G into a final graph G' with degree sequence S_{bound} by at most $\sum_{i=1}^{n} |S_G[i] - S_{bound}[i]|$ rotations. By Lemma 3.5.5 we know that $\sum_{i=1}^{n} |S_G[i] - S_{bound}[i]| \leq 2|OPT|$, hence we use at most 2 times the numbers of rotations of an optimal solution. In each transformation loop searching for the vertices v^+ and v^- can be done in time $O(n^2)$ and searching for an edge ab in time $O(m)$ (Lemma 2.5.2). Due to the modifications in each transformation loop, there can be at most $O(n^2)$ loops. Therefore the time complexity is bounded by $O(n^{2c} + (n^2 + m) \times n^2)$. Since $c \geq 4$ ($k \leq \frac{n}{4}$), $O(n^{2c} + n^4) \leq O(n^{2c})$.

Finally, since S_{bound} is k-anonymous, G' is a k-degree-anonymous graph. □

2.6 Polynomial cases

As follows from Section 3.5, the Min Anonymous-Edge-Rotation problem is NP-hard even for $k = \frac{n}{q}$ and $q \geq 3$ is a fixed constant where n is the order of an input graph. In this section we show that the problem can be solved in polynomial time on trees when $k = \Theta(n)$ or in case of any graph when $k = n$.

2.6.1 Trees

For a tree $T = (V, E)$ rooted at a vertex r, for any $v \in V$, $v \neq r$, $child(v)$ is a vertex that is a neighbor of v not on the path from r to v.

Lemma 2.6.1. *Let $T = (V, E)$ be a tree and v^-, v^+ vertices from V such that v^- is not a leaf and v^+ is not a universal vertex. Then using one rotation we can transform T into a tree T' such that $d_{T'}(v^-) = d_T(v^-) - 1$ and $d_{T'}(v^+) = d_T(v^+) + 1$.*

Proof. Let v^+ be the root of T. Since v^- is not a leaf, there exists a vertex $c \in child(v^-)$. Since T is a tree, $cv^+ \notin E$. Therefore we can define the rotation (cv^-, cv^+) (see Figure 3.13). Let T' be the graph obtained after such a rotation. Since there is no edge between the subtree of c and other vertices, T' is a tree. Moreover $d_{T'}(v^-) = d_T(v^-) - 1$ and $d_{T'}(v^+) = d_T(v^+) + 1$. □

Theorem 2.6.1. *The* Min Anonymous-Edge-Rotation *problem is polynomial-time solvable for any instance (T, k) where T is a tree of order n, $k \leq \frac{n}{4}$ and such that $c = \lfloor \frac{n}{k} \rfloor$ is a constant, hence $k = \Theta(n)$.*

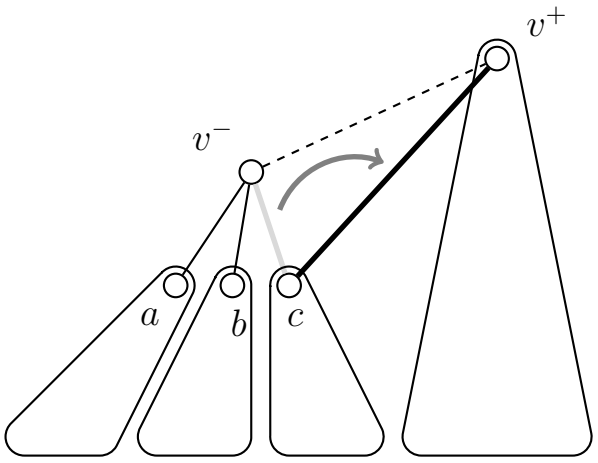


Figure 2.9: Transformation T to T'

Proof. Let T be a tree and $S_T = (d_1, d_2, \ldots, d_n)$ its degree sequence sorted in non-increasing order. For a degree sequence of a tree only the following conditions must hold $\sum_{i=1}^{n} d_i = 2(n-1)$ and $d_i \geq 1$ for all i, $1 \leq i \leq n$. Now based on S_T define a k-anonymous sequence S_{Tbound} as discussed in Section 3.5.

Let x and y be integers such that $S_T[x] > S_{Tbound}[x]$ and $S_T[y] < S_{Tbound}[y]$. Since S_{Tbound} corresponds to a tree, $S_{Tbound}[x] \geq 1$ then $S_T[x] > 1$ and thus v_x is not a leaf in T. Moreover since $S_{Tbound}[y] \leq n-1$, v_y is not a universal vertex in T.

By Lemma 3.5.1 there exists a tree T_1 such that $S_{T_1}[x] = S_T[x] - 1$ and $S_{T_1}[y] = S_T[y] + 1$. Repeat this operation until reaching a tree T' with the degree sequence S_{Tbound}. The runnning time of one operation is $O(n)$ and we repeat it $\frac{\sum_{i=1}^{n} |S_T[i] - S_{Tbound}[i]|}{2} \leq n^2$ times. Since S_{Tbound} is k-anonymous, T' is a k-degree-anonymous tree. Since we use $\frac{\sum_{i=1}^{n} |S_T[i] - S_{Tbound}[i]|}{2} \leq |OPT|$ rotations (Lemma 3.5.5), the algorithm is optimal. The total running time of the algorithm is bounded by $O(n^{2c} + n^2) = O(n^{2c})$, where $c = \lfloor \frac{n}{k} \rfloor$ is a constant greater than 4. □

2.6.2 One degree class, $k = n$

In this part we show that Min Anonymous-Edge-Rotation is polynomial-time solvable for instances where k coincides with the number of vertices of the graph, that means all vertices must be in the same degree class.

Lemma 2.6.2. *Let $G = (V, E)$ be a graph and $u, v \in V$. If $\mathcal{N}_G(u) \nsubseteq \mathcal{N}_G(v)$, then there is an edge rotation that leads to a graph G' such that $d_{G'}(u) = d_G(u) - 1$ and $d_{G'}(v) = d_G(v) + 1$.*

Proof. Since $\mathcal{N}_G(u) \nsubseteq \mathcal{N}_G(v)$, there exists $w \in V$ such that $uw \in E$ and $vw \notin E$. Then we can do the following edge rotation (uw, vw) and get the graph G' with $E' = (E \setminus \{uw\}) \cup \{vw\}$. □

Remark 2.6.1. *Let $G = (V, E)$ be a graph, $\forall u, v \in V$, if $d_G(u) > d_G(v)$, then there is an edge rotation that leads to a graph G' such that $d_{G'}(u) = d_G(u) - 1$ and $d_{G'}(v) = d_G(v) + 1$.*

Lemma 2.6.3. *Let (G, n) be an instance of* Min Anonymous-Edge-Rotation *where $G \in \mathcal{G}(n, m)$ for some positive integers m, n, and $\frac{2m}{n}$ is an integer. Then the optimum value of* Min Anonymous-Edge-Rotation *on (G, n) is $\frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$.*

Proof. As follows from Theorem 3.5.3, an instance (G, n) from $\mathcal{G}(n, m)$ is a feasible instance of Min Anonymous-Edge-Rotation if and only if $\frac{2m}{n}$ is an integer. Let suppose that (G, n) is such an instance. Obviously, if G is a regular graph, then the degree of each vertex must be $\frac{2m}{n}$.

If G is not a regular graph, then $\exists u, v \in V$ such that $d_G(u) > \frac{2m}{n}$ and $d_G(v) < \frac{2m}{n}$. By Remark 3.5.2, there is an edge rotation that leads to a graph G' such that $d_{G'}(u) = d_G(u) - 1$ and $d_{G'}(v) = d_G(v) + 1$. Then obviously at least $\frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$ rotations are necessary to have all the vertices of the same degree $\frac{2m}{n}$, therefore the optimum value of Min Anonymous-Edge-Rotation on the instance (G, n) is at least $\frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$.

Now suppose that the optimum value is r strictly less than $\frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$. Each rotation increases the degree of a vertex by one and decreases the degree of another vertex by one too. Obviously, each vertex w has to be involved in at least $|d_G(w) - 2m/n|$ edge rotations to reach the degree $\frac{2m}{n}$. Hence if there are $r < \frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$ edge rotations then in any graph G' obtained from G using r edge rotations there exists $w' \in V$ such that $d_{G'}(w') > \frac{2m}{n}$ or $d_{G'}(w') < \frac{2m}{n}$. □

Theorem 2.6.2. *The* Min Anonymous-Edge-Rotation *problem is polynomial-time solvable for instances (G, k) when $k = n$, where n is the order of the graph G.*

Proof. In case $k = n$, we are looking for a n-degree-anonymous graph with only one degree class, hence for a regular graph. Due to Theorem 3.5.3, we can easily decide whether (G, n) is a feasible instance of Min Anonymous-Edge-Rotation: if for $G \in \mathbf{G}(n, m)$ the fraction $\frac{2m}{n}$ is not an integer, (G, n) is not a feasible input.

For a feasible input (G, n), the result is based on Algorithm 1 and its correctness follows from Lemmas 3.5.2 and 2.6.3.

Obviously, the algorithm runs in polynomial time. □

2.7 Conclusion

We initiate the study of the complexity of Min Anonymous-Edge-Rotation problem in which the task is to transform a given graph into a k-degree anonymous graph using a minimum number of edge rotations. As we were able to prove NP-hardness in case where the number of vertices k in each degree class is $\Theta(n)$, further research could explore stronger hardness results or cases when k is a constant. A next research step could include relaxation of the condition on the number of edges in the presented 2-approximation algorithm as well as extension of the graph

Algorithm 1: Algorithm for $k = |V|$

Input : A graph $G = (V, E)$
Output: A sequence S of edge rotations if $\frac{2|E|}{|V|}$ is an integer
NO otherwise

$S = \emptyset$;
$d = \frac{2|E|}{|V|}$;
if *if d is not integer* **then**
  return NO ;
else
  while $\exists u, v \in V$ *such that* $d_G(u) < d$ *and* $d_G(v) > d$ **do**
    Let $w \in \mathcal{N}(v) \setminus \mathcal{N}(u)$;
    $E = E \setminus \{vw\}$;
    $E = E \cup \{uw\}$;
    $S = S \cup \{(wv, wu)\}$;
  end
end

classes in which the Min Anonymous-Edge-Rotation problem can be solved in polynomial time. As the problem does not have a solution for all graphs and all possible values of k, our initial feasibility study covers a large part of instances. The extensions of the results are still possible, in the sense of necessary and sufficient conditions.
Another interesting way to continue could be to study the same anonymization model but with the edge deletion/addition operator which consists in deleting an edge to reinject it at another place where an edge was missing. It has many similarities with edge rotation and their editing distances are related. The positive results seem easy to extend, however the hardness proof is too specific to rotation, a new one might be needed.

Chapter 3

Communities and Dense Graph Partitioning

In this chapter, we study the problem Max Dense Graph Partition of finding a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$, $k \geq 1$, of a given undirected graph G, such that the density of the partition, denoted by $d(\mathcal{P})$, is maximized. We consider a classical definition of the density of a subgraph induced by a subset S of vertices (see, for example, [21, 31]) given by the ratio between the number of edges and the number of vertices in S. The density of a partition is the sum of the densities of all its parts. Indeed, when the number of classes is given, the problem is a generalization of a partition into k cliques. We therefore address the problem Max Dense Graph Partition of finding a partition of maximum density, without fixing the number of classes of the partition.

3.1 Preliminaries

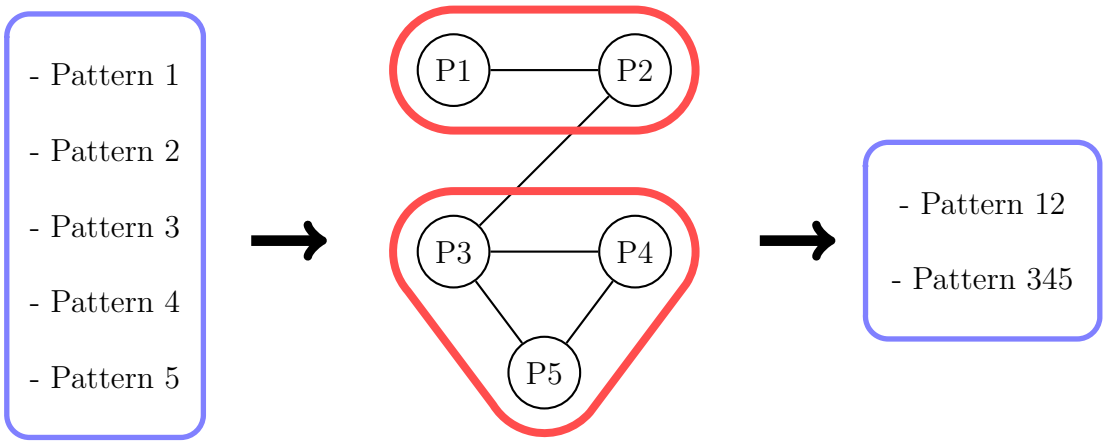
For this definition of density, there are several papers on finding the densest subgraph. This problem was shown solvable in polynomial-time by Goldberg [31] but if the size of the subgraph is a part on the input, the problem called k-Densest Subgraph becomes NP-hard even restricted to bipartite or chordal graphs [18]. The approximability of k-Densest Subgraph was also studied, see [41, 25, 9].

Medical Patterns

The problem was introduced by Darlay et al. [21] in the context of medical data analysis. During the PhD of Darlay [20], they faced a problem of community detection in medical data for clinical research. They used the logical analysis of data method [11] where a large set of patterns is generated (see Figure 3.14a), a pattern being the characteristics of patients having similar properties for the studied pathology. Their objective is to identify similar patterns to merge them in order to

The results of this section were published in [5] (arXiv version).

reduce their number. The idea is to partition the graph into communities of similar patterns (see Figure 3.14b). The patterns belonging to the same community will then be assimilated (see Figure 3.14c).



(a) Large pattern set generated

(b) Similarity graph and its partition into dense subgraphs

(c) Set of compressed patterns

Figure 3.1: Pattern compression process

Darlay et al. studied Max Dense Graph Partition, and its complement Min Sparse Graph Partition. They defined the sparsity of a partition $\mathcal{P}$ as $F(\mathcal{P}) = \frac{|\mathcal{P}|}{2} + d(\mathcal{P})$ and the problem Min Sparse Graph Partition as finding a partition of a given undirected graph G such that the sparsity of the partition is minimized. Observe that these two problems Max Dense Graph Partition and Min Sparse Graph Partition are dual in the sense that solving the first one on a graph G is the same as solving the second one on the complement of G. In [21] it is shown that both problems are NP-complete, and that there is no constant factor approximation for Min Sparse Graph Partition unless $P = NP$. Moreover, a polynomial-time algorithm for Max Dense Graph Partition on trees is given.

We point out that their proof of NP-completeness is a polynomial-time reduction from k-Coloring. By construction, the same reduction when starting from 3-Coloring on graphs of degree at most 4 (proved NP-complete in [28]) yields as instance of Max Dense Graph Partition a graph on n vertices and of minimum degree greater than $n-4n^{4/5}$. Thus it follows that Max Dense Graph Partition is NP-complete restricted to graphs of minimum degree $n - 4n^{4/5}$.

Coalitions and communities: Hedonic Game

Hedonic Game is a natural framework to study the formal aspect of coalition formation. A coalition can be seen as a community where the link between individuals is a common interest. This game models the formation of coalitions when players have preference over which group they belong to. Basically one wants to make

coalitions of very related people, which is close to the approach of partitioning into dense subgraphs. In an optimization problem related to the simple, symmetric and fractional Hedonic Game, Aziz et al. are interested in finding a partition that maximizes the utilitarian welfare [3], which is equivalent to the Dense Graph Partition problem. In this part we will first show that the problems are equivalent and then present their results.

The basic definition of an Hedonic Game is a set N of n elements and a set of complete and transitive relations $\succsim = (\succsim_1, \cdots, \succsim_n)$ that models the preferences of each player for the different coalitions (i.e., subsets of N).

The game is said to be *fractional* if for each player i a utility function $v_i : N \to \mathbb{R}$ is defined, which associates a value to the other players. Coalitions are seen as a set of distinct agents. To convert it into an Hedonic Game, an intermediate function is needed: let $v_i(S) = \frac{\sum_{j \in S} v_i(j)}{|S|}$ be the participation of the agent i in the coalition S. Using this function, an Hedonic Game is said *fractional* if for each player $i \in N$, for all coalitions $S, T \subseteq N$, $S \succsim_i T$ if and only if $v_i(S) \geq v_i(T)$.

A fractional Hedonic Game is said *symmetric* if $\forall i, j \in \{1, \cdots, n\}$, $v_i(j) = v_j(i)$ and *simple* if $\forall i, j \in \{1, \cdots, n\}$, $v_i \in \{0, 1\}$.

Lemma 3.1.1. *A simple, symmetric and fractional Hedonic Game $(N, \succsim)$ can be represented by a graph $G = (N, \{ij \subseteq N \mid v_i(j) = 1\})$.*

Proof. A simple fractional Hedonic Game is a set N of agents and a function v from $N \times N$ to $\{0, 1\}$. Since the game is symmetric $\forall i, j \in \{1, \cdots, n\}$, $v_i(j) = v_j(i)$, then it is isomorphic to a simple unweighted graph $G = (V, E) = (N, \{ij \subseteq N \mid v_i(j) = 1\})$. □

The *utilitarian welfare* of a partition is the sum of the utility of its agents. More formally, this means $\sum_{P \in \mathcal{P}} \sum_{i \in P} v_i(P)$.

Lemma 3.1.2. *The utilitarian welfare of a partition $\mathcal{P}$ of N is equal to two times the density of the same partition $\mathcal{P}$ in G.*

Proof. Since the game is fractional, the utilitarian welfare of a partition $\mathcal{P}$ is the sum of the utility of each agent in each coalition.

$$\sum_{P \in \mathcal{P}} \sum_{i \in P} v_i(P) = \sum_{P \in \mathcal{P}} \sum_{i \in P} \frac{\sum_{j \in P} v_i(j)}{|P|}$$

Since the game is simple and symmetric:

$$\begin{aligned} \sum_{i \in P} \frac{\sum_{j \in P} v_i(j)}{|P|} &= \sum_{i \in P} \frac{|\{j \in P \mid v_i(j) = 1\}|}{|P|} \\ &= \frac{2 \times |\{ij \in P \mid v_i(j) = 1\}|}{|P|} \\ &= \frac{2 \times |E(P)|}{|P|} \end{aligned}$$

Then we get:

$$\sum_{P\in\mathcal{P}}\sum_{i\in P} v_i(P) = \sum_{P\in\mathcal{P}} \frac{2\times|E(P)|}{|P|} = 2\times\sum_{P\in\mathcal{P}}\frac{|E(P)|}{|P|} = 2\times d(\mathcal{P})$$

□

We deduce that a partition which maximizes the utilitarian welfare in a simple symmetric fractional Hedonic Game $(N,\succsim)$ also maximizes the density of a graph $G=(N,\{ij\subseteq N \mid v_i(j)=1\})$ and reversely.

In [3] Aziz et al. proved that the problem is NP-complete even for 3-partite graphs. On the other hand they showed that a maximum matching provided a 2-approximation in the general case. There is no other work on this specific version of the Hedonic Game to our knowledge.

Our contributions. The following overview summarises the results achieved in this chapter concerning Max Dense Graph Partition.

- Max Dense Graph Partition is trivially solvable on graphs of maximum degree 2, we prove its NP-hardness for 3-regular (cubic) graphs.

- We establish that on bipartite complete graphs an optimal partition consists of one part, that is the whole graph. Moreover if the size of the two independent sets are relatively prime numbers then this optimal solution is unique. We use this result to show that Max Dense Graph Partition is $W[2]$-hard with respect to (an upper bound on) the number of clusters in an optimal solution on dense bipartite graphs. Our reduction is polynomial and hence in particular implies the NP-hardness of Max Dense Graph Partition on dense bipartite graphs.

- Max Dense Graph Partition is trivial on complete graphs since the optimal solution is the whole graph as one part of the partition. Moreover, as we previously explained, it is NP-hard on graphs of minimum degree $n-4n^{4/5}$. We show that for graphs of minimum degree $\geq n-3$, the problem is solvable in polynomial-time and any optimal solution has two parts. Moreover on $(n-4)$-regular graphs, the problem becomes NP-hard.

- We show that Max Dense Graph Partition admits a $(1+\varepsilon)$-approximation for any $\varepsilon>0$ on $(n-4)$-regular graphs, improving the 2-approximation on general graphs [3].

The chapter is organized as follows. Notations and formal definitions are given in the subsection below. The study of (dense) bipartite graphs is established in Section 3.5. Section 3.5 presents the results on cubic graphs. In Section 3.5 we study dense graphs. Some conclusions are given at the end of the chapter.

Notations

The density $d(G)$ of a graph $G = (V, E)$ is the ratio between the number of edges and the number of vertices in G, that is, $d(G) = \frac{|E|}{|V|}$. Moreover, for $S \subseteq V$, $d(S) = d(G[S]) = \frac{|E(S)|}{|S|}$. We use $\mathcal{P}$ to denote a partition of the set V of vertices of G, that is, $\mathcal{P} = \{V_1, \ldots, V_k\}$, where $\cup_{i=1}^{k} V_i = V$, and $V_i \cap V_j = \emptyset$ for each $i, j \in \{1, \ldots, k\}$. Then the density of the partition $\mathcal{P}$ of G is defined as $d(\mathcal{P}) = \sum_{i=1}^{k} d(G[V_i])$, where $G[V_i]$ is the subgraph of G induced by the subset V_i of vertices, that is, $G[V_i] = (V_i, E_i)$, $E_i = \{\{u, v\} : \{u, v\} \in E \wedge u, v \in V_i\}$.

We study the problem of finding a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$ of a given graph G, such that $k \geq 1$ and that, among all such partitions, $d(\mathcal{P})$ is maximized. We refer to this problem as Max Dense Graph Partition and we define its decision version as follows.

Dense Graph Partition
Input: An undirected graph $G = (V, E)$, a positive rational number r.
Question: Is there a partition $\mathcal{P}$ such that $d(\mathcal{P}) \geq r$?

We use some concepts from parameterized complexity and refer to [19, 22] for details of this terminology. A parameterized problem is a decision problem given together with a parameter, that is, an integer k depending on the instance. Such a parameterized problem is fixed-parameter tractable (fpt for short) if it can be solved in time $f(k) \cdot |I|^c$ for an instance I of size $|I|$ with parameter k, where f is a computable function and c is a constant. If a parameterized problem is hard for the complexity class W[2], it is unlikely (under certain complexity theoretic assumptions) to be fixed-parameter tractable. An fpt-reduction between two parameterized problems P and Q maps instances (I, k) of P to instances (I', k') of Q in time $f(k)|I|^{O(1)}$ for some computable functions f and g, such that $k' \leq g(k)$ and (I, k) is a yes-instance of P if and only if (I', k') is a yes-instance of Q. If problem P is W[2]-hard, then such an fpt-reduction shows that Q is W[2]-hard as well.

Given a maximization problem in NPO and an instance I of this problem, let OPT be the value of an optimal solution of I. For a function f, an algorithm is an $f(|I|)$-approximation if, for every instance I of the problem, it returns a solution with a value S such that $OPT \leq f(|I|) \times S$. Moreover if the algorithm runs in polynomial-time in $|I|$, then this algorithm gives a polynomial-time $f(|I|)$-approximation. When $f = 1 + \varepsilon$, for any $\varepsilon > 0$, the problem admits a polynomial-time approximation scheme. When the running time of an approximation scheme is of the form $O(g(1/\varepsilon)poly(|I|)$ the problem has an efficient polynomial-time approximation scheme (eptas).

Before we start studying specific graph classes, we observe the following helpful structural properties that hold for Dense Graph Partition on general graphs.

Remark. *We can assume that for any optimal partition $\mathcal{P}$ and for any part $P_i \in \mathcal{P}$, $G[P_i]$ is connected, since otherwise turning each connected component into its own part does not decrease the density.*

Lemma 3.1.3. *Among all partitions of G into $t \geq 2$ parts, those where the parts correspond to complete graphs, if there exists such, have the largest density.*

Proof. Consider a partition of G into t parts $\{V_1, \ldots, V_t\}$ of size $n_1, \ldots, n_t$. If $G[V_i]$ has o_i missing edges for any $1 \leq i \leq t$, then the density of this partition is $\frac{n-t}{2} - \frac{o_1}{n_1} - \ldots - \frac{o_t}{n_t}$.

Consider a partition of G into t parts of size $n'_1, \ldots, n'_t$ such that each part induces a complete graph for any $1 \leq i \leq t$. Then the density of this partition is $\frac{n-t}{2}$ and thus it is larger than the density of any partition in t parts where at least one edge is missing inside $G[V_i]$ for some $1 \leq i \leq t$. □

A direct consequence of this is the following.

Lemma 3.1.4. *Let $G = (V, E)$ be a graph and $\mathcal{P}$ be any partition of V. Then $d(\mathcal{P}) \leq \frac{|V|}{2} - \frac{|\mathcal{P}|}{2}$.*

3.2 Dense Bipartite Graphs

In this section we show that Max Dense Graph Partition has a trivial solution on complete bipartite graphs. Moreover, using this result we show that the problem is NP-hard on dense bipartite graphs and even $W[2]$-hard with respect to the number of clusters in an optimal solution as parameter.

In the first part, we consider a complete bipartite graph $G_{n,m}$ with the two subsets that are independent sets of size n and m and we first prove the following result.

Lemma 3.2.1. *The density $d(G_{n,m})$ of a complete bipartite graph $G_{n,m}$ is greater than or equal to the density $d(\mathcal{P})$ of any partition $\mathcal{P}$ of $G_{n,m}$.*

Proof. The density of the complete bipartite graph $G_{n,m} = (A, B, E)$, with $|A| = n, |B| = m$ is given by $d(G_{n,m}) = \frac{nm}{n+m}$

It suffices to show that $d(G_{n,m})$ is greater than or equal to the density of any partition $\mathcal{P} = \{V_1, V_2\}$ that splits the set of vertices into exactly 2 nonempty subsets. Indeed, if this holds and we have a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$ where $k \geq 3$, we can show recursively that $d(G_{n,m}) \geq d(G[V_1]) + d(G[V_2 \cup \cdots \cup V_k]) \geq \cdots \geq d(G[V_1]) + \cdots + d(G[V_k])$.

We first consider a partition $\mathcal{P}_1 = \{V_1, V_2\}$ where $A \subseteq V_1$. Without loss of generality we may assume that $V_2 = B \setminus V_1$ contains m_2 vertices from B. Then

$$d(\mathcal{P}_1) = \frac{n(m - m_2)}{n + m - m_2} + 0 \leq \frac{nm}{n+m}$$

Now, consider a partition $\mathcal{P}_1 = \{V_1, V_2\}$ such that each of the graphs $G[V_i]$ contains at least one edge, so let $G[V_1] = G_{n_1,m_1}$ with $0 < n_1 < n$ and $0 < m_1 < m$. Then $G[V_2] = G_{n-n_1,m-m_1}$ and

$$d(\mathcal{P}_1) = \frac{n_1 m_1}{n_1 + m_1} + \frac{(n - n_1)(m - m_1)}{n + m - n_1 - m_1} = \frac{nm(n_1 + m_1) - mn_1^2 - nm_1^2}{(n + m - n_1 - m_1)(n_1 + m_1)},$$

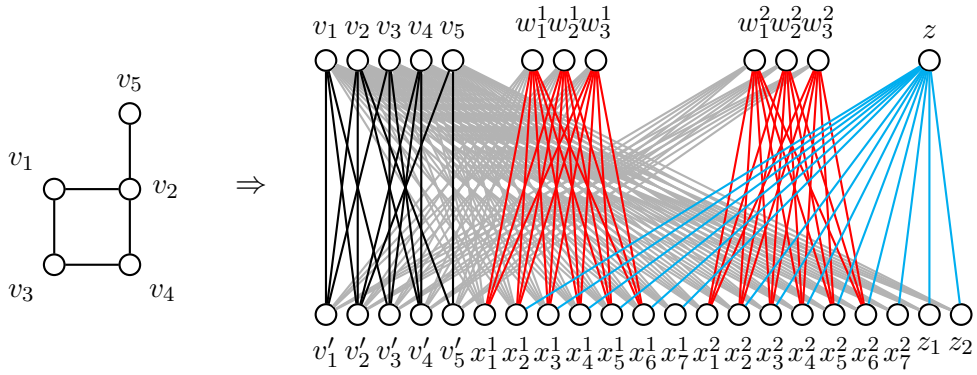


Figure 3.2: A graph G, instance of Dominating Set and the bipartite graph G' obtained from G, for $k = 2$ and $n = 5$.

which yields

$$d(G_{n,m}) - d(\mathcal{P}_1) = \frac{(nm_1 - mn_1)^2}{(n + m - n_1 - m_1)(n_1 + m_1)(n + m)} \geq 0$$

□

It follows that an optimal solution of any complete bipartite graph is the whole graph. From the calculations in the previous proof, we can inductively deduce the following result.

Corollary 3.2.1. *For any complete bipartite graph $G = (A, B, E)$ with $|A| = n$ and $|B| = m$, a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$ of $A \cup B$ satisfies $d(\mathcal{P}) = \frac{nm}{n+m}$ if and only if $G[V_i] = G_{n_i,m_i}$ with $n_i \neq 0$ and $m_i \neq 0$ and $\frac{n_i}{m_i} = \frac{n}{m}$ for all $i \in \{1, \ldots, k\}$.*

Consequently, for any complete bipartite graph $G_{n,m}$, if n and m are relatively prime the only optimal solution of $G_{n,m}$ is the whole graph. Otherwise, several optimal solutions exist and are characterized exactly by Corollary 3.5.2.

In the second part of this section, we study the role of the number of sets in an optimum solution for Dense Graph Partition. We specifically consider parameterization in the following sense. To formally give a parameter as input, we consider instances of the form $((G, r), k)$ for the decision problem: Is there a partition $\mathcal{P}$ of the vertices of G into at most k sets with $d(\mathcal{P}) \geq r$?

Theorem 3.2.1. *Dense Graph Partition parameterized by (an upper bound on) the number of clusters in an optimal solution is W[2]-hard on dense bipartite graphs (i.e. $|E| = \Theta(n^2)$).*

Proof. We give a reduction from Dominating Set. Given a graph $G = (V, E)$ with $V = \{v_1, \ldots, v_n\}$ and an integer $k \geq 1$ as instance of Dominating Set, we construct a bipartite graph $G' = (V_1, V_2, E')$ as input for Dense Graph Partition as follows:

- $V_1 = V \cup \{w_i^j : 1 \leq i \leq n - k,\ 1 \leq j \leq k\} \cup \{z\}$

- $V_2 = \{v'_1, \ldots, v'_n\} \cup \{x^j_r \colon 1 \leq r \leq N,\ 1 \leq j \leq k\} \cup \{z_i \colon 1 \leq i \leq N - n\}$ where $N \in \mathbb{N}$ is chosen as follows. Let $c \in \mathbb{N}$ be the smallest integer such that $c(n-k+1)-1 > n$ (note that $1 \leq c \leq n$) and define $N = c(n-k+1)-1$. For this choice of N it follows that the greatest common divisor of N and $n-k+1$ is 1, and $n < N \leq 2n$.

- $E' = E_d \cup E_{wx} \cup E_c \cup E_z$ with
  $E_d = \{\{v_i, v'_j\} \colon \{v_i, v_j\} \in E\} \cup \{\{v_i, v'_i\} \colon 1 \leq i \leq n\}$,
  $E_{wx} = \{\{w^j_i, x^j_r\} \colon 1 \leq i \leq n-k,\ 1 \leq r \leq N-1, 1 \leq j \leq k\}$,
  $E_c = \{\{w^j_i, v'_s\} \colon 1 \leq i \leq n-k, 1 \leq j \leq k,\ 1 \leq s \leq n\} \cup \{\{v_s, x^j_r\} \colon 1 \leq s \leq n,\ 1 \leq r \leq N,\ 1 \leq j \leq k\}$ and
  $E_z = \{\{z, z_j\} \colon 1 \leq j \leq N-n\} \cup \{\{z, x^j_r\} \colon 2 \leq r \leq N, 1 \leq j \leq k\} \cup \{\{v_i, z_j\} \colon 1 \leq i \leq n,\ 1 \leq j \leq N-n\}$

Notice that G' is a bipartite graph with $|V_1| = n+1+k(n-k)$ and $|V_2| = (k+1)N$.

We show that there exists a dominating set of cardinality at most k in G if and only if there exists a partition $\mathcal{P}$ of G' with $d(\mathcal{P}) = (k+1)d(G_{n-k+1,N})$.

Suppose there exists a dominating set D in G with $|D| = k$. Let $D = \{v_{i_1}, \ldots, v_{i_k}\}$ and $N'(v_{i_j}) = N_G[v_{i_j}] \backslash (D \cup N_G(\{v_{i_1}, \ldots, v_{i_{j-1}}\})$. Define the partition $\mathcal{P} = \{P_1, \ldots, P_{k+1}\}$ by:
$P_j = \{v_{i_j}\} \cup \{v'_{i_j}\} \cup \{v'_r \colon v_r \in N'(v_{i_j})\} \cup \{w^j_r \colon 1 \leq r \leq n-k\} \cup \{x^j_r \colon 1 \leq r \leq N - |N'(v_{i_j})|\}$ for $1 \leq j \leq k$ and $P_{k+1} = V_1 \cup V_2 \setminus (\cup_{j=1}^k P_j)$. It is not hard to see that the vertices in P_j induce a complete bipartite graph $G_{n-k+1,N}$ for each j. Thus $d(\mathcal{P}) = (k+1)d(G_{n-k+1,N})$.

Conversely, let $\mathcal{P}$ be a partition of G' of density $(k+1)d(G_{n-k+1,N})$. Thus, Corollary 3.5.2 implies that the vertices for each set $P \in \mathcal{P}$ induce a complete bipartite graph $G_{r,s}$ such that $\frac{r}{s} = \frac{|V_1|}{|V_2|} = \frac{k(n-k)+n+1}{(k+1)N} = \frac{n-k+1}{N}$. Since the greatest common divisor of $n-k+1$ and N is one, this yields $r \geq n-k+1$ and $s \geq N$ and especially $\mathcal{P}$ can contain at most $k+1$ sets.

For all w^j_i and w^t_ℓ, if $j \neq t$, w^j_i and w^t_ℓ have n common neighbours, and since $n < N$ there is no part $P \in \mathcal{P}$ such that $w^j_i, w^t_\ell \in P$. Moreover, for all i, j, w^j_i and z have $N-1$ common neighbours so they cannot be in the same $P \in \mathcal{P}$. Hence, there are exactly $k+1$ parts in $\mathcal{P}$ that are complete bipartite graphs $G_{n-k+1,N}$.

For all $1 \leq j \leq k$, denote by P_j the set containing the vertices w^j_i for all $1 \leq i \leq n-k$ and P_z the part containing z. To reach the cardinality exactly $n-k+1$, $P_j \cap V_1$ has to contain exactly one vertex from V for each $1 \leq j \leq k$. Further, since for any i, v'_i is not adjacent to z, $V' \subseteq \cup_{j=1}^k P_j$. Moreover, each P_j contains exactly one vertex of V. As each $P \in \mathcal{P}$ induces a complete bipartite graph in G', $D = V \cap \cup_{j=1}^k P_j$ is a set of size k, such that each vertex in V' is adjacent to at least one vertex in D, so we deduce that D is a dominating set of size k in G.

At last, in case of a yes-instance of Dominating Set, there exists an optimum solution with $k+1$ sets for Dense Graph Partition on G'. With parameter $k' = k+1$, the instance $((G', (k+1)d(G_{n-k+1,N})), k')$ describes an fpt-reduction from Dominating Set parameterized by solution size to Dense Graph Partition parameterized by an upper bound on the number of sets. Since Dominating Set is W[2]-hard, this reduction implies that Dense Graph Partition is also W[2]-hard.

We extend the construction of the proof to create from G' a dense bipartite graph $G'' = (V'', E'')$ by adding four sets of vertices $V_1^u, V_1^d, V_2^u, V_2^d$ with $|V_1^u| = |V_1^d| = kn|V_1| = kn(k(n-k)+n+1)$ and $|V_2^u| = |V_2^d| = kn|V_2| = knN(k+1)$. Further, we add edges to turn the pairs (V_1^u, V_2^u), (V_1^d, V_2^d), (V_1^u, V_1), and (V_2^d, V_2) each into complete bipartite graphs. Observe that with this construction G'' has $|V''| = (2kn+1)(k(n-k)+n+1) + (2kn+1)N(k+1) < 10k^2n^2$ vertices and that all vertices have degree at least $kn|V_1| \geq \frac{1}{2}k^2n^2 \in \Theta(|V''|)$ (Note that Dominating Set is trivial when $k \geq \frac{n}{2}$.).

We claim that there exists a partition $\mathcal{P}'$ of G'' with $d(\mathcal{P}') = (k+1)d(G_{n-k+1,N}) + 2kn(k+1)d(G_{n-k+1,N})$ if and only if there exists a dominating set of size k for G. Corollary 3.5.2 again implies that this density for G'' can only be achieved by a partition into complete bipartite graphs $G_{r,s}$ with $\frac{r}{s} = \frac{(2kn+1)(k(n-k)+n+1)}{(2kn+1)N(k+1)} = \frac{n-k+1}{N}$. The vertices in V_1^d are only adjacent to vertices in V_2^d, and the vertices in V_2^u are only adjacent to vertices in V_1^u. Clustering these in a ratio $\frac{r}{s}$ results in clusters containing exactly all newly added vertices, and this can be done with just two sets in total. What remains is to cluster the graph G' into complete bipartite graphs $G_{r,s}$ such that $\frac{r}{s} = \frac{|V_1|}{|V_2|} = \frac{k(n-k)+n+1}{(k+1)N} = \frac{n-k+1}{N}$ as before.

At last, in case of a yes-instance of Dominating Set, there exists an optimum solution with $k+3$ sets for Dense Graph Partition on G''. With parameter $k' = k+3$, the instance $((G'', (2kn+1)(k+1)d(G_{n-k+1,N})), k')$ describes an fpt-reduction from Dominating Set parameterized by solution size to Dense Graph Partition parameterized by an upper bound on the number of sets, which shows the claimed W[2]-hardness. □

3.3 Cubic Graphs

We show that the problem Dense Graph Partition is NP-complete even for cubic graphs by giving a reduction from Exact Cover By 3-Sets where each element appears in exactly 3 sets, denoted Restricted Exact Cover By 3-Sets, known to be NP-hard by [32].

Restricted Exact Cover By 3-Sets (RX3C)
Input: A set X of elements with $\|X\| = 3q$ and a collection C of 3-element subsets of X where each element appears in exactly 3 sets.
Question: Does C contain an exact cover for X, i.e. a subcollection $C' \subseteq C$ such that every element occurs in exactly one member of C' ?

Before describing the reduction, we give useful notions for utility.

Definition 3.3.1. *For $S \subseteq V$, the utility of a vertex $v \in S$ is defined by $u_S(v) = \frac{d(S)}{|S|}$, and the utility of S is defined by $u(S) = u_S(v)$ for any $v \in S$. For a partition $\mathcal{P} = \{V_1, \ldots, V_k\}$, the utility of a vertex v in $\mathcal{P}$ is defined by $u_{\mathcal{P}}(v) = u_{V_i}(v)$ with i such that $v \in V_i$.*

Considering these definitions, we can remark that:

- For any subset $S \subseteq V$, and $v, w \in S$, $u_S(v) = u_S(w)$.

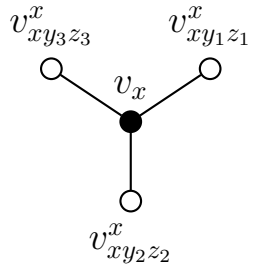


Figure 3.3: Subgraph containing one vertex of type 1, v_x, and its neighbors in G

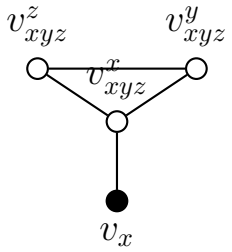


Figure 3.4: Subgraph containing one vertex of type 1, v_x, and three of type 2

- If $S = \{v\}$ then $u_S(v) = 0$.
- For any partition $\mathcal{P}$ of G, $\sum_{V_i \in \mathcal{P}} d(V_i) = \sum_{v \in V} u_{\mathcal{P}}(v)$.

The following definition gives the construction to reduce RX3C to Dense Graph Partition .

Definition 3.3.2. *Let $I = (X, C)$ be an instance of RX3C. We define the construction σ transforming the instance I into the graph $G := \sigma(I)$ where $G = (V, E)$ as follows:*

- *for each element $x \in X$, we add the vertex v_x in V (called vertices of type 1 or black vertices).*
- *for each subset of the collection $\{x, y, z\} \in C$, we add the vertices v^x_{xyz}, v^y_{xyz}, v^z_{xyz} in V (called vertices of type 2 or white vertices).*
- *we add the edges $\{v^x_{xyz}, v^y_{xyz}\}$, $\{v^x_{xyz}, v^z_{xyz}\}$ and $\{v^y_{xyz}, v^z_{xyz}\}$ to E*
- *we add the edges $\{v^x_{xyz}, v_x\}$, $\{v^y_{xyz}, v_y\}$ and $\{v^z_{xyz}, v_z\}$ to E*

Notice that G is a cubic graph on $|X|$ vertices of type 1 and $3|X|$ vertices of type 2.

Case distinction on the subgraphs in $\sigma(I)$ shows:

Lemma 3.3.1. *For any subset $S \subseteq V$ of the vertices of the graph $\sigma(I)$, the only subgraphs $G[S]$ with $u(S) \geq \frac{1}{4}$ are:*

- *a triangle where all the vertices are of type 2 and then $u(S) = \frac{1}{3}$.*
- *an edge between two type 2 vertices or between two vertices of different types and then $u(S) = \frac{1}{4}$.*
- *the subgraph described in Figure 3.4 and then $u(S) = \frac{1}{4}$.*

Proof. Let $S \subseteq V$ such that $u(S) \geq \frac{1}{4}$. We show in the following that there are exactly three possible subgraphs $G[S]$ such that $u(S) \geq \frac{1}{4}$. First, observe that by its construction G does not contain C_4 as subgraph, since there are no two vertices $u, v \in V$ that have more than one common neighbor. Note that this also implies that G is diamond-free (i.e. without $K_4 - e$ subgraphs).

As G is cubic, $|E(G[S])| \leq \frac{3}{2}|S|$ and so $d(S) \leq \frac{3}{2}|S| \cdot \frac{1}{|S|} = \frac{3}{2}$. Since $\frac{1}{4} \leq u(S) \leq \frac{3}{2|S|}$ then $|S| \leq 6$. We study the five following cases :

- Case $|S| = 6$: Since $u(S) = \frac{|E(S)|}{6^2} \geq \frac{1}{4}$, we have $|E(S)| \geq 9$. Since $G[S]$ cannot be cubic (G is connected and $|V| > 6$), a subgraph with $|S| = 6$ and $|E(S)| \geq 9$ does not exist.

- Case $|S| = 5$: Since $u(S) = \frac{|E(S)|}{5^2} \geq \frac{1}{4}$, we have $|E(S)| \geq 7$. Assuming such a subgraph with $|S| = 5$ and $|E(S)| \geq 7$, $G[S]$ must contain a diamond or a clique of size 4. Since $G = \sigma(I)$ is diamond-free and G is cubic, such a subgraph does not exist.

- Case $|S| = 4$: Since $u(S) = \frac{|E(S)|}{4^2} \geq \frac{1}{4}$, we have $|E(S)| \geq 4$. Since G is diamond-free, S is not a K_4 or a C_4 the only possibility for $G[S]$ is the subgraph described in Figure 3.4.

- Case $|S| = 3$: Since $u(S) = \frac{|E(S)|}{3^2} \geq \frac{1}{4}$, we have $|E(S)| \geq 3$ and thus S is a triangle where all the vertices are of type 2 and $u(S) = \frac{1}{3}$.

- Case $|S| = 2$: Since $u(S) = \frac{|E(S)|}{2^2} \geq \frac{1}{4}$, we have $|E(S)| \geq 1$ and thus S is an edge between two type 2 vertices or between two vertices of different types and $u(S) = \frac{1}{4}$.

□

Remark 3.3.1. *For any subset $S \subseteq V$ of the vertices of the graph $\sigma(I)$, if v is of type 2 then $u_S(v) \leq \frac{1}{3}$, otherwise $u_S(v) \leq \frac{1}{4}$.*

With these observations about the construction of $\sigma(I)$, it can be shown that $I = (X, C)$ is a yes-instance of RX3C if and only if $I' = \sigma(I)$ is a yes-instance of Dense Graph Partition which yields the following.

Theorem 3.3.1. *Dense Graph Partition is NP-complete on cubic graphs.*

Proof. Let $I = (X, C)$ be an instance of RX3C and consider the following instance I' of Dense Graph Partition on the graph $G = \sigma(I)$ and $d = \frac{7|X|}{6}$. We claim that $I = (X, C)$ is a yes-instance of RX3C if and only if $I' = (G, d)$ is a yes-instance of Dense Graph Partition .

Let $C' \subseteq C$ be an exact cover for X of size $\frac{|X|}{3}$. Consider the following partition $\mathcal{P}$ with $\frac{5|X|}{3}$ parts : for any $c \in C'$, $c = \{x, y, z\}$, we define three parts of size 2, $\{v_x, v^x_{xyz}\}$, $\{v_y, v^y_{xyz}\}$, $\{v_z, v^z_{xyz}\}$ and for any $c \notin C'$, $c = \{x, y, z\}$, we define the following part of size 3, $\{v^x_{xyz}, v^y_{xyz}, v^z_{xyz}\}$. Since C' is an exact cover, $\mathcal{P}$ is a partition and its density is $\frac{3}{2} \cdot \frac{|X|}{3} + \frac{2}{3}|X| = \frac{7}{6}|X|$.

Let $\mathcal{P}'$ be a partition of G of density $d(\mathcal{P}') = \frac{7}{6}|X|$. Firstly, we show that $\mathcal{P}'$ has necessarily the following shape: $\frac{2|X|}{3}$ parts of size 3 containing only vertices of type 2 forming a triangle in G and $|X|$ parts of size 2 containing one vertex of type 1 and one of type 2 adjacent in G (see Figures 3.3 and 3.4). From Remark 3.1, we can consider that for every part $P_i \in \mathcal{P}'$, $G[P_i]$ is connected.

We prove in the following that since $d(\mathcal{P}') = \frac{7|X|}{6}$ then there are at least $\frac{2|X|}{3}$ parts in $\mathcal{P}'$ corresponding to triangles in G. Assume by contradiction that $\mathcal{P}'$ has $\frac{2|X|}{3} - \ell$ triangles, with $\ell > 0$. Since G has $4|X|$ vertices, there are $2|X| + 3\ell$ vertices that do not belong to a part in $\mathcal{P}'$ that corresponds to a triangle in G. By Lemma 3.5.7 the utility of these last vertices is smaller than or equal to $\frac{1}{4}$. Then the density of $\mathcal{P}'$ is

$$d(\mathcal{P}') \leq \frac{2|X|}{3} - \ell + (2|X| + 3\ell) \cdot \frac{1}{4} = \frac{7|X|}{6} - \frac{\ell}{4} < \frac{7|X|}{6}$$

This contradicts the choice of $\mathcal{P}'$ such that $d(\mathcal{P}') = \frac{7|X|}{6}$, hence there are at least $\frac{2|X|}{3}$ triangles in $\mathcal{P}'$.

Now, we will prove that there are at most $\frac{2|X|}{3}$ parts in $\mathcal{P}'$ corresponding to triangles in G. Assume by contradiction that $\mathcal{P}'$ has $\frac{2|X|}{3} + \ell$ triangles, with $\ell > 0$. Since there are $3|X|$ vertices of type 2 and among these vertices $3 \cdot (\frac{2|X|}{3} + \ell)$ belong to a triangle then $|X| - 3\ell$ vertices of type 2 do not belong to a triangle. But each neighbour of a vertex v_x of type 1 is of type 2, so if the utility of v_x is positive, then there exists a vertex of type 2, v^x_{xyz}, neighbour of v_x, that is in the same part as v_x and v^x_{xyz} does not belong to a triangle. Moreover, as all type 1 vertices have no common neighbours, for each type 1 vertex with positive utility, there is a type 2 vertex that is not in a triangle. Since there are at most $|X| - 3\ell$ type 2 vertices that do not belong to a triangle, there are at most $|X| - 3\ell$ type 1 vertices with positive utility. Then the density of $\mathcal{P}'$ is at most

$$d(\mathcal{P}') \leq \frac{2|X|}{3} + \ell + \frac{|X| - 3\ell}{4} + \frac{|X| - 3\ell}{4} \leq \frac{7|X|}{6} - \frac{\ell}{2} < \frac{7|X|}{6}$$

This contradicts the choice of $\mathcal{P}'$ such that $d(\mathcal{P}') = \frac{7|X|}{6}$, and then there are exactly $\frac{2|X|}{3}$ triangles in $\mathcal{P}'$.

We will show now that $d(\mathcal{P}') = \frac{7|X|}{6}$ implies that all type 1 vertices are in a part that is a matching with a type 2 vertex. There are $|X|$ type 1 vertices and $|X|$ type 2 vertices that are not in some triangle in $\mathcal{P}'$. Since there are exactly $\frac{2|X|}{3}$ parts in $\mathcal{P}'$ forming a triangle and the utility of each other vertex is smaller than or equal to $\frac{1}{4}$, to reach a density of $\frac{7|X|}{6}$ it is necessary that each of the $2|X|$ vertices outside the parts that are triangles has a utility of exactly $\frac{1}{4}$. To reach this utility, by Lemma 3.5.7 there are two possibilities, the graph described in Figure 3.4 and an edge. Since there are exactly $|X|$ vertices of type 1 and $|X|$ vertices of type 2 outside the triangles in $\mathcal{P}'$, and vertices of type 1 only have neighbors of type 2, the only possibility for all these vertices to have utility $\frac{1}{4}$ is if each type 1 vertex is matched with one type 2 vertex.

Consider now the following subcollection $C'' \subseteq C$: for each triple v^x_{xyz},v^y_{xyz},v^z_{xyz} that does not belong to a triangle, we add the set $\{x, y, z\}$ to C''. The subcollection C'' is a cover since each type 1 vertex is a neighbour of one of these vertices and it

is an exact cover since there are exactly $\frac{|X|}{3}$ 3-element subsets that do not belong to a triangle. □

3.4 Highly Dense Graphs

In this section we consider graphs $G = (V, E)$ on n vertices such that G can be viewed as $G = K_n - H$ where H is a graph of small maximum degree. The edges of H are called *missing edges* in G. We first consider graphs $G = (V, E)$ on n vertices such that $\delta(G) \geq n - 3$, that is $G = K_n - H$ where H has $\Delta(H) = 2$ and has $q \leq n$ edges and show that Max Dense Graph Partition is solvable in polynomial-time on these graphs.

Remark 3.4.1. *Section 3.5 deals with graphs such that $m = \Theta(n^2)$ and this section deals with graphs such that $\delta \geq n - c$ with $c \in \{1, \ldots, 4\}$.*

Lemma 3.4.1. *For any graph G on n vertices such that $\delta(G) \geq n - 3$, its density $d(G)$ is greater than or equal to the density of any partition $\mathcal{P}$ of G into $t \geq 3$ parts.*

Proof. The density of G is given by $d(G) = \frac{\frac{n(n-1)}{2} - q}{n} = \frac{n-1}{2} - \frac{q}{n}$. From Lemma 3.1.3, among all partitions of G into $t \geq 3$ parts, those where the parts correspond to complete graphs have the largest density. The density of such a partition into t parts of size $n_1, \ldots, n_t$ is $\frac{n-t}{2}$. Thus, the density of G is at least as large as the density of this last partition since $t \geq 3$ and $q \leq n$ (note here that a graph with minimum degree $n - 3$ has at most n non-edges). □

Remark that in the proof of the previous lemma when $q = n$ and $t = 3$, the density of a partition in 3 parts corresponding to complete subgraphs and the density of the entire graph are the same. This previous lemma implies that for any graph G such that $\delta(G) \geq n - 3$, there exists a partition into one or two parts of maximum density.

Lemma 3.4.2. *For any graph G on n vertices such that $\delta(G) \geq n - 3$, in any partition into two parts of G, the number of missing edges inside the two parts is at least o, where o is the number of odd cycles defined by the missing edges of G.*

Proof. Let C be and odd cycle of missing edges in G. Since C is not bipartite, there is no partition $\{V_1, V_2\}$ of V such that all the edges of C have one endpoint in V_1 and one endpoint in V_2. Hence, for any partition $\{V_1, V_2\}$ at least one of the missing edges from C is inside $G[V_1] \cup G[V_2]$. □

Lemma 3.4.3. *Among all partitions into 2 parts of fixed size containing x missing edges, the one containing all missing edges in the largest part has the best density.*

Proof. Consider two partitions $\{V_1, V_2\}$ and $\{V_1', V_2'\}$ such that $|V_1| = |V_1'| = n_1$ and $|V_2| = |V_2'| = n_2$ with $n_1 \leq n_2$ and $G[V_1]$ (resp. $G[V_2]$) containing x_1 (resp. x_2) missing edges and $G[V_1']$ (resp. $G[V_2']$) containing 0 (resp. $x = x_1 + x_2$) missing edges.

$d(\{V_1, V_2\}) = \frac{n-2}{2} - \frac{x_1}{n_1} - \frac{x_2}{n_2}$
$d(\{V_1', V_2'\}) = \frac{n-2}{2} - \frac{x}{n_2}$
Since $x = x_1 + x_2$ and $n_1 \leq n_2$, we have $d(\{V_1, V_2\}) \leq d(\{V_1', V_2'\})$. □

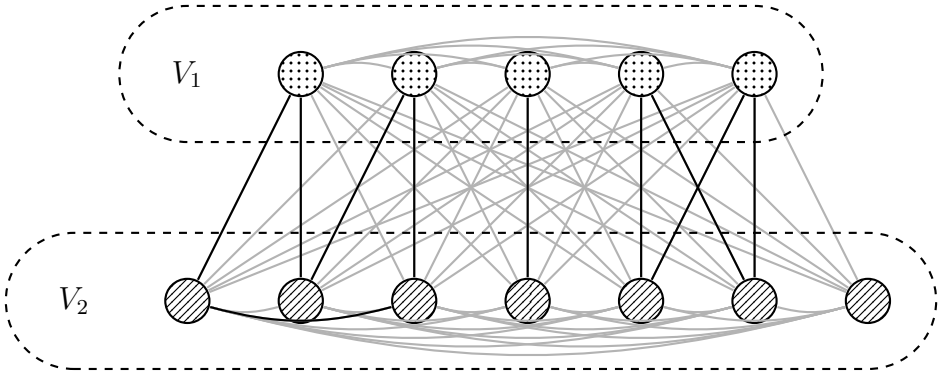


Figure 3.5: Construction of V_1 and V_2 in Theorem 3.5.11

Lemma 3.4.4. *Among all partitions into 2 parts containing 0 (resp. x) missing edges in the smaller (resp. larger) part, the one with a maximum number of vertices in the largest part has the best density.*

Proof. Consider two partitions $\{V_1, V_2\}$ and $\{V_1', V_2'\}$ such that $|V_1| = n_1$, $|V_2| = n_2$ with $n_1 \leq n_2$ and $|V_1'| = n_1'$, $|V_2'| = n_2'$ with $n_1' \leq n_2'$ and $G[V_1]$ (resp. $G[V_2]$) containing 0 (resp. x) missing edges and $G[V_1']$ (resp. $G[V_2']$) containing 0 (resp. x) missing edges. Moreover suppose $n_2 \leq n_2'$.

$d(\{V_1, V_2\}) = \frac{n-2}{2} - \frac{x}{n_2}$

$d(\{V_1', V_2'\}) = \frac{n-2}{2} - \frac{x}{n_2'}$

Since $n_2 \leq n_2'$, we have $d(\{V_1, V_2\}) \leq d(\{V_1', V_2'\})$. □

Theorem 3.4.1. Max Dense Graph Partition *is solvable in polynomial-time on graphs G with n vertices and $\delta(G) \geq n - 3$.*

Proof. We define a partition $\{V_1, V_2\}$ where V_1 (resp. V_2) contains vertices of color 1 (resp. 2). An example is given in Figure 3.5. Each vertex of degree $n-1$ has color 2. The graph H of missing edges contains paths or cycles. The vertices on paths or cycles with an even number of vertices are colored alternating by 1 and 2. The vertices on paths or cycles with an odd number of vertices are colored alternating by 1 and 2 but starting with color 2. Thus cycles of odd size have two adjacent vertices of color 2. The partition $\{V_1, V_2\}$ defined above is such that it contains o missing edges in V_2 and $|V_2|$ is maximized among all such partitions. Its density is equal to $\frac{n-2}{2} - \frac{o}{n_2}$, where $n_2 = |V_2|$. Denote by d_{n-1} the number of vertices of G of degree $n-1$ and by p_o the number of paths with an odd number of vertices (even length) among the missing edges. Thus $n_2 = \frac{1}{2}(n + d_{n-1} + p_o + o)$. We claim that there is no partition into two parts that has a higher density.

By Lemma 3.4.2, any partition into two sets contains at least o missing edges inside the two parts. By construction we have maximized the number of vertices in the part with the missing edges among all partitions with the minimum number o of missing edges, i.e., there is no partition into two parts $\{V_1', V_2'\}$ with o missing edges all contained in V_2' and $|V_2'| > |V_2|$. Hence, by Lemma 3.4.3 and 3.4.4, it remains to show that any partition $\{V_1', V_2'\}$ with $n_2' = |V_2|$ such that $n_2' = n_2 + y$ with $o + x > o$ missing edges have a smaller density than $\{V_1, V_2\}$.

By definition of the partition $\{V_1, V_2\}$, it follows that $|E(H)| = 2n_1 - r + o$, where r is the number of paths of odd length in H. For the partition $\{V_1', V_2'\}$, it follows that $|E(H)| \leq 2(n_1 - y) - r_1 + (o + x)$, for some $r_1 \geq r - x$ (number of vertices in V_1' adjacent to only one edge in H). Observe that all non-edges have to either be among the $o + x$ missing edges in the partition or in the cut between V_1' and V_2'. In the cut between V_1' and V_2', each vertex in V_1' is adjacent to at most two such edges, and further every path of odd length either results in a vertex in V_1' adjacent to only one edge in $E(H)$ (r_1) or in a missing edge in V_2', hence $r_1 \geq r - x$. These inequalities imply that $y \leq x$, and hence the density of $\{V_1', V_2'\}$ is at most $\frac{n-2}{2} - \frac{o+x}{n_2+y} \leq \frac{n-2}{2} - \frac{o+y}{n_2+y} \leq \frac{n-2}{2} - \frac{o}{n_2}$. Note that the last inequality follows from $o \leq n_2$, which simply holds since H is of degree at most 2. □

In the rest of the section we consider graphs $G = (V, E)$ on n vertices, $(n-4)$-regular, that is $G = K_n - H$ where H is a cubic graph. We show that Dense Graph Partition is NP-hard on $(n-4)$-regular graphs, by showing a reduction from unCut on cubic graphs, that is the complement of Max Cut. This last problem on cubic graphs was proved NP-hard and even not polynomial-time 1.003-approximable, unless P=NP [8].

Min UnCut
Input: A graph $G = (V, E)$, an integer k.
Question: Does G contain a partition of V into two parts A, B such that the number of edges with both endpoints in the same part is at most k?

Since we did not find a reference for the following result in the literature we propose a short proof.

Lemma 3.4.5. *Let $G = (V, E)$ be a cubic graph. There exists a partition $\{A, B\}$ of G with a cut of size at least $|V|$ and it can be found in polynomial-time.*

Proof. Let $\mathcal{P} = \{A, B\}$ be a partition of V. Consider the following operation: if there is a vertex $v \in A$ (resp. B) with at least two neighbours in A (resp. B) then $A = A \setminus \{v\}$ (resp. $B = B \setminus \{v\}$) and $B = B \cup \{v\}$ (resp. $A = A \cup \{v\}$). Since the graph is cubic, this operation increases the number of edges between A and B by at least one. Since the number of edges is finite, we can repeat this operation until we obtain a partition $P' = \{A', B'\}$ with no vertex $v \in A'$ (resp. B') with at least two neighbours in A' (resp. B'). Since the graph is cubic, if every vertex in A' (resp. B') has at most one neighbour in A', then it has at least two neighbours in B' (resp. A'). Consequently P' has a cut of size at least $\frac{2(|A'|+|B'|)}{2} = |V|$. □

Definition 3.4.1. *Let $I = (G, k)$ be an instance of unCut where $G = (V, E)$ is a cubic graph. We define the construction σ transforming the graph G into the graph $G' := (V', E') = \sigma(G)$ (see Figure 3.6) as follows:*

- *let $G_0 = (V_0, E_0)$ be the union of $\frac{n^2-n}{6}$ copies of $K_{3,3}$ (see remark below). Thus G_0 is a cubic bipartite graph with $n^2 - n$ vertices and V_0 is the union of two independent sets L, R such that $|L| = |R|$.*

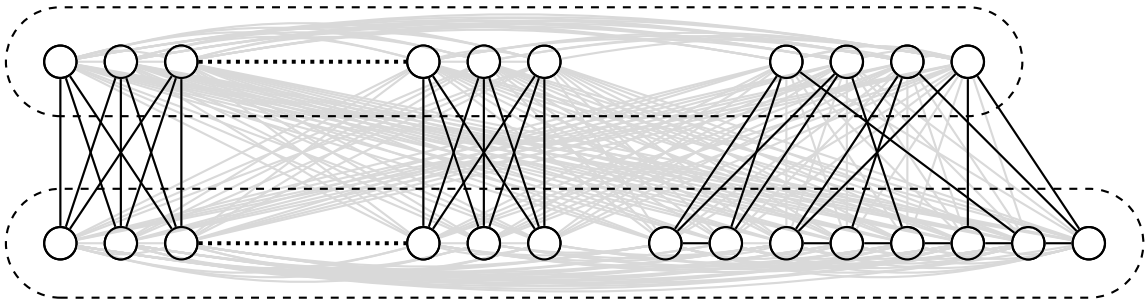


Figure 3.6: The construction of G' in Definition 3.4.1

- *let* $G_1 = (V \cup V_0, E \cup E_0)$.
- *let* $G' = \overline{G_1}$.

Remark 3.4.2. *Note that we can assume that the number of vertices of a cubic graph G is a multiple of 6. Since G is cubic, n is a multiple of 2. If n is not a multiple of 3, we consider the instance I_{triple} defined as follows: G_{triple} is the union of 3 copies of G and $k_{triple} = 3k$, and thus in the new instance I_{triple} the graph has $3n$ vertices.*

Let $n = |V|$, $m = |E|$, $n' = |V'|$ and $m' = |E'|$. Remark that $n' = n^2$, and G' is a $(n' - 4)$-regular graph.

Since we reduce from Min UnCut on cubic graphs, we use the following straight forward observation on any partition in such graphs.

Lemma 3.4.6. *For any cubic graph G and any $\{A, B\}$ partition of V, we have $|A| + \frac{2}{3} \cdot |E(B)| = |B| + \frac{2}{3} \cdot |E(A)|$, where $E(A)$, resp. $E(B)$, is the set of edges with both endpoints in A, resp B.*

Proof. Since the graph G is cubic, $|E(A, B)| = 3 \cdot |A| - 2 \cdot |E(A)| = 3 \cdot |B| - 2 \cdot |E(B)|$. We can deduce that $|A| + \frac{2}{3} \cdot |E(B)| = |B| + \frac{2}{3} \cdot |E(A)|$ □

Theorem 3.4.2. Dense Graph Partition *is NP-complete on $(n - 4)$-regular graphs with n vertices.*

Proof. Let $I = (G = (V, E), k)$ be an instance of unCut , where G is a cubic graph. Consider the following instance I' of Dense Graph Partition on the graph $G' = \sigma(G)$ and $d = \frac{n^2}{2} - 1 - \frac{2k}{n^2}$. We claim that $I = (G, k)$ is a yes-instance of unCut if and only if $I' = (G', d)$ is a yes-instance of Dense Graph Partition .

Let $\{A, B\}$ be a partition of V whose uncut value is at most k. Since $V_0 = L \cup R$, where L, R are independent sets in G_0 such that $|L| = |R|$, the sets L, R form two cliques of the same size in G'. Let $A' = A \cup L$ and $B' = B \cup R$ and $\mathcal{P} = \{A', B'\}$ be a partition of G'.

Let $M_{A'}$ and $M_{B'}$ be the set of missing edges in $G'[A']$ and $G'[B']$, respectively. Due to the construction of G', there is no missing edge between A and L and between

B and R. Thus all missing edges are inside $G'[A \cup B]$, i.e. $|M_{A'}| + |M_{B'}| = k$. Thus, the density of the partition $\mathcal{P}$ is:

$$d(\mathcal{P}) = \frac{|A'|-1}{2} - \frac{|M_{A'}|}{|A'|} + \frac{|B'|-1}{2} - \frac{|M_{B'}|}{|B'|} = \frac{n^2-2}{2} - \frac{|M_{A'}|}{|A'|} - \frac{|M_{B'}|}{|B'|}$$

We will prove in the following that $d(\mathcal{P}) \geq d = \frac{n^2}{2} - 1 - \frac{2k}{n^2}$ that is equivalent to proving that $\frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|} \leq \frac{2(|M_{A'}|+|M_{B'}|)}{|A'|+|B'|}$.

Consider the difference

$$\frac{2(|M_{A'}| + |M_{B'}|)}{|A'| + |B'|} - \left(\frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|}\right)$$

$$= \frac{1}{|A'| + |B'|}\left(2|M_{A'}| + 2|M_{B'}| - \frac{|A'| + |B'|}{|A'|}|M_{A'}| - \frac{|A'| + |B'|}{|B'|}|M_{B'}|\right)$$

$$= \frac{1}{|A'| + |B'|}\frac{1}{|A'|}\frac{1}{|B'|}(|A'||B'||M_{A'}| + |A'||B'||M_{B'}| - |B'|^2|M_{A'}| - |A'|^2|M_{B'}|)$$

$$= \frac{1}{|A'| + |B'|}\frac{1}{|A'|}\frac{1}{|B'|}(|A'| - |B'|)(|B'||M_{A'}| - |A'||M_{B'}|)$$

Wlog we can consider that $|A'| \geq |B'|$, that implies $|B'| \leq \frac{n^2}{2}$. From Lemma 3.4.6 for the cubic graph G_1 and partition $\{A', B'\}$, we have $|A'| + \frac{2}{3} \cdot |M_{B'}| = |B'| + \frac{2}{3} \cdot |M_{A'}|$. Using that $|A'| = n^2 - |B'|$ and $|M_{A'}| = k - |M_{B'}|$, we have $n^2 - |B'| + \frac{2}{3} \cdot |M_{B'}| = |B'| + \frac{2}{3} \cdot (k - |M_{B'}|)$ and thus $|M_{B'}| = \frac{3}{4}(2|B'| + \frac{2}{3}k - n^2)$.

Thus,

$$|B'||M_{A'}| - |A'||M_{B'}| = |B'|(k - |M_{B'}|) - (n^2 - |B'|)|M_{B'}| = |B'|k - n^2|M_{B'}|$$

$$= |B'|k - n^2\frac{3}{4}\left(2|B'| + \frac{2}{3}k - n^2\right) = \left(|B'| - \frac{n^2}{2}\right)\left(k - \frac{3n^2}{2}\right)$$

Since $|B'| \leq \frac{n^2}{2}$ and $k \leq \frac{n}{2} \leq \frac{3n^2}{2}$ we can conclude that

$$\frac{2(|M_{A'}| + |M_{B'}|)}{|A'| + |B'|} - \left(\frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|}\right) \geq 0$$

Thus, the partition $\mathcal{P} = \{A', B'\}$ has density $d(\mathcal{P}) \geq d = \frac{n^2}{2} - 1 - \frac{2k}{n^2}$.

Let $\mathcal{P}'$ be a partition of G' of density $d(\mathcal{P}') \geq d = \frac{n^2-2}{2} - \frac{2k}{n^2}$. We will prove that $\mathcal{P}'$ has exactly two parts A' and B' such that $A = A' \cap V$ and $B = B' \cap V$ is a partition of G whose uncut value is at most k.

Suppose that $|\mathcal{P}'| \geq 3$. Then, using Lemma 3.1.4, we have $d(\mathcal{P}') \leq \frac{n^2-|\mathcal{P}'|}{2} \leq \frac{n^2-3}{2} = \frac{n^2-2}{2} - \frac{1}{2}$. Since $k \leq \frac{n}{2}$ and $n \geq 6$ then $\frac{2k}{n^2} < \frac{1}{2}$. Then $d(\mathcal{P}') < \frac{n^2-2}{2} - \frac{2k}{n^2} = d$ which is a contradiction. Then $|\mathcal{P}'| < 3$.

Suppose that $|\mathcal{P}'| = 1$. Since G' is (n^2-4)-regular, its density is $d(\mathcal{P}') = \frac{n^2-1}{2} - \frac{3}{2} = \frac{n^2-2}{2} - 1 < \frac{n^2-2}{2} - \frac{2k}{n^2} = d$ which is a contradiction. Then $|\mathcal{P}'| > 1$. We conclude that $|\mathcal{P}| = 2$.

Let A' and B' be the two parts of $\mathcal{P}$. Let $M_{A'}$, resp. $M_{B'}$, be the set of missing edges in $G'[A']$, resp. $G'[B']$. Remark that if $|M_{A'}|+|M_{B'}| \leq k$ then $|M_A|+|M_B| \leq k$ and then there is a cut of size at least k between A and B in G. What it remains to prove is that $|M_{A'}| + |M_{B'}| \leq k$.

As a first step we will show the following inequality we need later $\frac{|M_{A'}|+|M_{B'}|}{\frac{n^2}{2}+\frac{|M_{A'}|+|M_{B'}|}{3}} \leq \frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|}$. In order to prove this, we consider the following difference

$$\frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|} - \frac{|M_{A'}| + |M_{B'}|}{\frac{|A'|+|B'|}{2} + \frac{|M_{A'}|+|M_{B'}|}{3}}$$

By removing the denominator we get

$$|M_{A'}||B'|\left(\frac{|A'|+|B'|}{2} + \frac{|M_{A'}|+|M_{B'}|}{3}\right) + |M_{B'}||A'|\left(\frac{|A'|+|B'|}{2} + \frac{|M_{A'}|+|M_{B'}|}{3}\right)$$

$$-(|M_{A'}| + |M_{B'}|)|A'||B'| =$$

$$= |M_{A'}||B'|\left(\frac{|B'|}{2} + \frac{|M_{A'}|}{3} + \frac{|M_{B'}|}{3} - \frac{|A'|}{2}\right)+|M_{B'}||A'|\left(\frac{|A'|}{2} + \frac{|M_{B'}|}{3} + \frac{|M_{A'}|}{3} - \frac{|B'|}{2}\right) =$$

From Lemma 3.4.6 for the cubic graph G_1 and partition $\{A', B'\}$, we have $|A'| + \frac{2}{3}|M_{B'}| = |B'| + \frac{2}{3}|M_{A'}|$, which implies that $\frac{|A'|}{2} = \frac{|B'|}{2} + \frac{|M_{A'}|}{3} - \frac{|M_{B'}|}{3}$ and $\frac{|B'|}{2} = \frac{|A'|}{2} + \frac{|M_{B'}|}{3} - \frac{|M_{A'}|}{3}$ and then we get

$$= |M_{A'}||B'|\left(\frac{|B'|}{2} + \frac{|M_{A'}|}{3} + \frac{|M_{B'}|}{3} - \left(\frac{|B'|}{2} + \frac{|M_{A'}|}{3} - \frac{|M_{B'}|}{3}\right)\right)$$

$$+|M_{B'}||A'|\left(\frac{|A'|}{2} + \frac{|M_{B'}|}{3} + \frac{|M_{A'}|}{3} - \left(\frac{|A'|}{2} + \frac{|M_{B'}|}{3} - \frac{|M_{A'}|}{3}\right)\right) =$$

$$= |M_{A'}||B'|\frac{2|M_{B'}|}{3} + |M_{B'}||A'|\frac{2|M_{A'}|}{3}$$

Since $|M_{A'}|$, $|M_{B'}|$, $|A'|$ and $|B'|$ are positive integers $\frac{|M_{A'}|}{|A'|}+\frac{|M_{B'}|}{|B'|}-\frac{|M_{A'}|+|M_{B'}|}{\frac{n^2}{2}+\frac{|M_{A'}|+|M_{B'}|}{3}} \geq 0$. We conclude that $\frac{|M_{A'}|+|M_{B'}|}{\frac{n^2}{2}+\frac{|M_{A'}|+|M_{B'}|}{3}} \leq \frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|}$.

Finally, we show that $|M_{A'}| + |M_{B'}| \leq k$ using the previous inequality. Let $x = |M_{A'}|+|M_{B'}|$. In order to finalize the proof, we suppose that $x > k$ and we will arrive at a contradiction, that is $d(\mathcal{P}') < d$. Consider the following difference

$$d - d(\mathcal{P}') = \frac{n^2-2}{2} - \frac{2k}{n^2} - \left(\frac{n^2-2}{2} - \frac{|M_{A'}|}{|A'|} - \frac{|M_{B'}|}{|B'|}\right) = \frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|} - \frac{2k}{n^2}$$

Since $\frac{x}{\frac{n^2}{2}+\frac{x}{3}} \leq \frac{|M_{A'}|}{|A'|} + \frac{|M_{B'}|}{|B'|}$

$$d - d(\mathcal{P}') \geq \frac{x}{\frac{n^2}{2}+\frac{x}{3}} - \frac{2k}{n^2} = \frac{x \cdot n^2 - k \cdot n^2 - \frac{2x \cdot k}{3}}{(\frac{n^2}{2}+\frac{x}{3}) \cdot n^2}$$

Since x and k are integers, $x \geq k+1$, and by removing the denominator, we get

$$\geq (k+1) \cdot (n^2 - \frac{2}{3} \cdot k) - k \cdot n^2 = n^2 - \frac{2}{3} \cdot k^2 - \frac{2}{3} \cdot k$$

Since $k \leq \frac{n}{2}$ it follows that $n^2 - \frac{2}{3} \cdot k^2 - \frac{2}{3} \cdot k > 0$. This finally gives $d(\mathcal{P}') < d$, a contradiction to the choice of $\mathcal{P}'$ as partition with density at least d, and we hence conclude that $|M_{A'}| + |M_{B'}| \leq k$.

Overall, it follows that if $d(\mathcal{P}') \geq \frac{n^2-2}{2} - \frac{2k}{n^2}$ then there is a partition $\{A, B\}$ with an uncut of size at most k. □

At the end of this section we show that a partition into three cliques provides a good approximation of the problem.

Lemma 3.4.7. *Let $G = (V, E)$ be a $(n-4)$-regular graph and $\mathcal{P}$ any partition of V. Then $d(\mathcal{P}) \leq \frac{n}{2} - 1$.*

Proof. If $|\mathcal{P}| = 1$ then $d(\mathcal{P}) = \frac{n-4}{2}$. Suppose that $|\mathcal{P}| \geq 2$, the density is maximized when for every $P \in \mathcal{P}$, $G[P]$ is a clique. Then $d(\mathcal{P}) = \sum_{P \in \mathcal{P}} \frac{|P|-1}{2} \leq \frac{n}{2} - 1$. □

Theorem 3.4.3. *There is an efficient polynomial-time approximation scheme for* Max Dense Graph Partition *on $(n-4)$-regular graphs.*

Proof. Let $I = G$ be a graph on n vertices and $(n-4)$-regular, instance of Max Dense Graph Partition. Let $\overline{G}$ be the complementary graph of G. By Brooks' theorem, we know that there is a 3-coloration of $\overline{G}$ that can be found in polynomial-time [38].

We establish in the following an eptas. Given $\varepsilon > 0$, consider two cases. If $n \geq 3 + \frac{1}{\varepsilon}$, then let $\mathcal{P}$ be a partition, that corresponds to a 3-coloration of $\overline{G}$, such that each part is a clique in G. Then $d(\mathcal{P}) = \frac{n}{2} - \frac{3}{2}$. By Lemma 3.5.9 we know that $opt(I) \leq \frac{n}{2} - 1$. Thus $d(\mathcal{P}) \geq \frac{n/2-1}{1+\varepsilon} \geq \frac{opt(I)}{1+\varepsilon}$.

Otherwise, that is $n < 3 + \frac{1}{\varepsilon}$, enumerate all the partitions of G and consider the best one. Since the number of partitions of G is the Bell number of order $|V| = n$, B_n, and $B_n \leq n^n$, we get an optimal solution in time $O((1/\varepsilon)^{O(1/\varepsilon)})$. □

3.5 Conclusion

We continued the study of Max Dense Graph Partition initiated by Darlay et al. in [21] and continued by Aziz et al. in [3]. The problem consists in identifying the communities/parts/coalitions that constitute a graph through the density criterion. As Darlay et al. have showed that the problem is polynomial-time solvable on trees, an interesting future research direction could be to study the complexity of the problem parameterized by the minimum feedback edge set. On the other hand there is a 2-approximation of the problem, we have even shown that it is possible to do better in dense graphs. For the moment there are no inapproximability results, further research could explore this area to give a better understanding of the border of approximation. Still in dense graphs, we should see if there is an XP algorithm (parameterized by the number of parts) or not in order to finish the work started

on the W[2]-hardness. The fact that the density objective function is not an integer implies problems of an algorithmic order but also for the proofs. As we have seen in the proofs of this section, simply comparing two quantities can be a difficult task. It would therefore seem important to find an algorithmic and proof strategy that can overcome this difficulty.

Conclusion

This thesis presents results on two original problems, a degree editing problem where the solution is an (ordered) sequence of edges related to graph anonymization and a partitioning problem optimizing a rational objective function. We have developed algorithmic techniques adapted to the problems and we have identified cases in which the problems became polynomial time solvable although they are intractable in the general case.

Concerning the anonymization problem, two open perspectives are proposed. First, extending the results on k-degree-anonymization on more constrained models. Precisely, new advanced models were proposed (like i-hop-anonymization, k-neighbourhood-anonymization or k-automorphism-anonymization, see [14]). Each of them has its particular interest. However, no theoretical study has been dedicated to classify the complexity of the anonymization problem on these models. Second, such questions deserve to be tackled as a multidisciplinary question, i.e, adapting statistical and probabilistic methods may lead to solve interesting questions. There are intermediate models proposing a tradeoff between the accuracy and the data loss among the k-anonymization model and the ϵ-differential privacy model (see [43]). In these models one could lower the required epsilon of differential privacy, which would have the effect of impacting the data utility loss. To compensate for the weakening caused by the decrease of the ϵ, one could k-anonymize, which is generally less destructive for the data (with a low value of k). Finally, establishing a theoretical link between these two models may, first validate the experimental results obtained in [37], second lead to interesting equivalent guarantees (such that if a graph is k-degree-anonymous then it is $f(k)$-differentially private) on some restricted graph classes.

Regarding density, many variations could be proposed to match other community detection needs. It would be interesting to be able to require a minimum part density (or a minimum vertex utility) to guarantee a minimal quality for every community. The problem is that we need an adaptive bound to tackle the following problem: a high bound in a graph with sparse components can lead to the nonexistence of solutions. On an other hand, a lower bound on the size of parts might be interesting, with the current formulation it is quite possible to return communities of a single isolated vertex in a regular graph.

Bibliography

[1] Charu C. Aggarwal and Philip S. Yu. *Privacy-preserving data mining: models and algorithms*. Springer Science & Business Media, 2008.

[2] Richard D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973.

[3] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Julián Mestre, and Hanjo Taubig. Welfare maximization in fractional hedonic games. In *24th International Joint Conference on Artificial Intelligence*, pages 461–467. IEEE, 2015.

[4] Cristina Bazgan, Robert Bredereck, Sepp Hartung, André Nichterlein, and Gerhard J. Woeginger. Finding large degree-anonymous subgraphs is hard. *Theoretical Computer Science*, 622:90–110, 2016.

[5] Cristina Bazgan, Katrin Casel, and Pierre Cazals. Dense graph partitioning on sparse and dense graphs. *arXiv:2107.13282*, 2021.

[6] Cristina Bazgan, Pierre Cazals, and Janka Chlebíková. How to get a degree-anonymous graph using minimum number of edge rotations. In *International Conference on Combinatorial Optimization and Applications*, pages 242–256. Springer, 2020.

[7] Cristina Bazgan, Pierre Cazals, and Janka Chlebíková. Degree-anonymization using edge rotations. *Theoretical Computer Science*, 873:1–15, 2021.

[8] Piotr Berman and Marek Karpinski. On some tighter inapproximability results. In *26th International Colloquium on Automata, Languages, and Programming*, pages 200–209. Springer, 1999.

[9] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings of the 42nd Symposium on Theory of Computing*, pages 201–210. ACM, 2010.

[10] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: theory and experiment*, 2008(10):P10008, 2008.

[11] Endre Boros, Peter L Hammer, Toshihide Ibaraki, Alexander Kogan, Eddy Mayoraz, and Ilya Muchnik. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, 12(2):292–306, 2000.

[12] Robert Bredereck, Vincent Froese, Sepp Hartung, André Nichterlein, Rolf Niedermeier, and Nimrod Talmon. The complexity of degree anonymization by vertex addition. *Theoretical Computer Science*, 607:16–34, 2015.

[13] Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. k-degree anonymity and edge selection: improving data utility in large networks. *Knowledge and Information Systems*, 50(2):447–474, 2017.

[14] Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. A survey of graph-modification techniques for privacy-preserving on networks. *Artificial Intelligence Review*, 47(3):341–366, 2017.

[15] Jordi Casas-Roma, Julián Salas, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. k-degree anonymity on directed networks. *Knowledge and Information Systems*, 61(3):1743–1768, 2019.

[16] Gary Chartrand, Farrokh Saba, and Hung Bin Zou. Edge rotations and distance between graphs. *Časopis pro pěstování matematiky*, 110(1):87–91, 1985.

[17] Edgar F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[18] Derek G. Corneil and Yehoshua Perl. Clustering and domination in perfect graphs. *Discrete Applied Mathematics*, 9(1):27–39, 1984.

[19] Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015.

[20] Julien Darlay. *Analyse combinatoire de données: structures et optimisation*. PhD thesis, Université de Grenoble, 2011.

[21] Julien Darlay, Nadia Brauner, and Julien Moncel. Dense and sparse graph partition. *Discrete Applied Mathematics*, 160(16-17):2389–2396, 2012.

[22] Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013.

[23] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[24] Paul Erdős and Tibor Gallai. Gráfok eloírt fokú pontokkal (graphs with points of prescribed degrees, in Hungarian). *Mat. Lapok*, 11:264–274, 1961.

[25] Uriel Feige, Guy Kortsarz, and David Peleg. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[26] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[27] Michael R. Garey and David S. Johnson. *Computers and Intractability.* Freeman, San Francisco, 1979.

[28] Michael R. Garey, David S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1(3):237–267, 1976.

[29] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12, 2005.

[30] Michelle Girvan and Mark E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[31] Andrew V. Goldberg. *Finding a maximum density subgraph.* University of California Berkeley, 1984.

[32] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[33] Louis S. Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. I. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962.

[34] Sepp Hartung, Christian Komusiewicz, André Nichterlein, and Ondřej Suchỳ. On structural parameterizations for the 2-club problem. *Discrete Applied Mathematics*, 185:79–92, 2015.

[35] Sepp Hartung, André Nichterlein, Rolf Niedermeier, and Ondřej Suchỳ. A refined complexity analysis of degree anonymization in graphs. *Information and Computation*, 243:249–262, 2015.

[36] Michael Hay, Gerome Miklau, David Jensen, Philipp Weis, and Siddharth Srivastava. Anonymizing social networks. *Computer Science Department Faculty Publication Series, article 180*, 2007.

[37] Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. (k,ϵ)-anonymity: k-anonymity with ϵ-differential privacy. *arXiv:1710.01615*, 2017.

[38] Howard J. Karloff. An NC algorithm for Brooks' theorem. *Theoretical Computer Science*, 68(1):89–103, 1989.

[39] Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.

[40] Brian W. Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307, 1970.

[41] Subhash Khot. Ruling out PTAS for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM J. Comput.*, 36(4):1025–1071, 2006.

[42] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[43] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[44] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD 2008*, pages 1–34, 2008.

[45] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[46] Duncan R. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2):169–190, 1950.

[47] Robert D. Luce and Albert D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[48] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge*, 3(2):151–166, 2013.

[49] Spiros Mancoridis, Brian S. Mitchell, Chris Rorres, Yih F. Chen, and Emden R. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *6th International Workshop on Program Comprehension*, pages 45–52. IEEE, 1998.

[50] Alex Pothen. Graph partitioning algorithms with applications to scientific computing. In *Parallel Numerical Algorithms*, pages 323–368. Springer, 1997.

[51] Veronica Red, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[52] Julián Salas and Vicenç Torra. Graphic sequences, distances and k-degree anonymity. *Discrete Applied Mathematics*, 188:25–31, 2015.

[53] John Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[54] Jordi Soria-Comas. Improving data utility in differential privacy and k-anonymity. *arXiv:1307.0966*, 2013.

[55] Latanya Sweeney. Uniqueness of simple demographics in the us population. *LIDAP-WP4*, 2000.

[56] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[57] Ioan Tomescu. Problems in combinatorics and graph theory. *Wiley-Interscience Series in Discrete Mathematics*, pages 212–213, 1961.

[58] Jinbao Wang, Zhipeng Cai, Yingshu Li, Donghua Yang, Ji Li, and Hong Gao. Protecting query privacy with differentially private k-anonymity in location-based services. *Personal and Ubiquitous Computing*, 22(3):453–469, 2018.

[59] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and Applications*. Cambridge University Press, 1994.

[60] Xiaowei Ying and Xintao Wu. Randomizing social networks: a spectrum preserving approach. In *SIAM International Conference on Data Mining*, pages 739–750, 2008.

[61] Elena Zheleva and Lise Getoor. Preserving the privacy of sensitive relationships in graph data. In *International Workshop on Privacy, Security, and Trust in KDD*, pages 153–171. Springer, 2007.

Résumé en français

Introduction

La quantité de données collectées ne cesse d'augmenter, que ce soit en termes de quantité de données, de leur variété ou du nombre d'acteurs qui les collectent. Cette croissance est rendue possible par l'accessibilité toujours plus grande à un terminal informatique, au réseau Internet et par la diminution des coûts de stockage des serveurs. Selon le cabinet d'études IDC, le volume de stockage mondial passerait de 33 zettaoctets en 2018 à 175 zettaoctets en 2025. Selon la même source, 75% de ces données sont détenues par des entreprises. La taille du marché européen des données est estimée à 184 milliards d'euros en 2020 et devrait atteindre entre 200 et 300 milliards d'euros en 2025, selon le portail européen des données.

Les détenteurs de données ont du mal à divulguer ces informations sans compromettre la vie privée des individus. A travers le GDPR (General Data Protection Regulation), on observe qu'une dynamique de renforcement de la législation autour de la détention de données personnelles est en cours. Au final, dans de nombreux cas, la survie de ces bases de données dépendra de la capacité du détenteur à produire des données anonymes pour permettre leur exploitation sans nuire à autrui.

Actuellement, les détenteurs de données opèrent individuellement, ignorant la possibilité de recoupement avec d'autres données. Une pratique courante appelée pseudonymisation (encouragée par le GDPR) consiste à traiter les données de manière à ce qu'il soit impossible de les attribuer à un individu sans l'aide de données supplémentaires. Dans la plupart des cas, les données restantes peuvent être utilisées pour ré-identifier les individus en faisant correspondre les données avec d'autres bases de données existantes. Un exemple connu est l'expérience de Latanya Sweeney : Elle a d'abord montré que la combinaison du code postal, du sexe et de la date de naissance était unique pour 87% des Américains [55]. Dans un deuxième temps, voir Figure 3.7, elle a utilisé deux bases de données publiques, la "liste d'inscription des électeurs de Cambridge Massachusetts" contenant le quadruplet : Nom, adresse, code postal, sexe des habitants de Cambridge et une seconde, produite par la "Group Insurance Commission", responsable de l'achat des assurances pour les employés de l'État, contenant les dossiers médicaux ainsi que le code postal, le sexe et la date de naissance de chaque individu bien que par leur nom. Par un simple recoupement, elle a pu identifier le dossier médical du gouverneur de cet État. Il est donc difficile de garantir un anonymat suffisant tout en permettant l'exploitation des données.

Pour contribuer à ce défi, nous proposons d'identifier les modèles qui garantissent une anonymisation "adéquate" dans les bases de données de graphes. Nous chercherons ensuite à concevoir des algorithmes qui transforment les données de

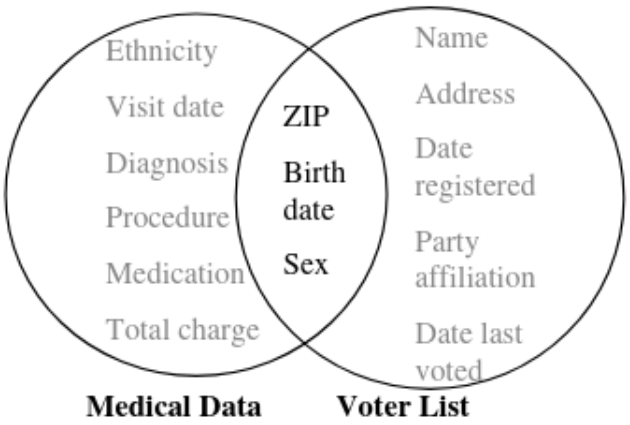


Figure 3.7: Data cross-checking

manière à respecter ces modèles tout en minimisant à la fois les perturbations induites et le temps de calcul. Plus précisément, notre objectif est de concevoir des algorithmes efficaces d'anonymisation des graphes avec des garanties de performance et de donner un aperçu de la difficulté de calcul. Nous essayons de nous concentrer sur des paramètres aussi proches que possible de la situation pratique, en termes de niveau d'anonymat, de structure de graphe et de perte de données.

D'autre part, nous nous intéressons à la structure de la base de données des graphes. Les réseaux réels ne sont pas des graphes aléatoires : ils présentent de fortes particularités et sont généralement des graphes épars avec peu de sommets de haut degré. De plus, la distribution des arêtes est aussi localement inhomogène avec des ensembles de sommets fortement connectés et une très faible interaction entre ces groupes (voir exemple dans Figure 2). Cette organisation spécifique des réseaux réels est appelée partition de communauté [30].

Plus concrètement, les sociétés présentent une grande variété de groupes sociaux : relationnels (famille, travail, amitié), géographiques (village, ville, pays), culturels (langue, hobbies, musique), etc. D'autre part, l'internet a permis la création de communautés qui n'existaient pas auparavant (forums, réseaux sociaux, jeux vidéo).

Si la notion de communauté peut sembler intuitive, il est plus difficile d'en donner une définition formelle. En effet, il n'existe pas de définition universelle et souvent la définition dépendra de l'application visée. Cependant, il existe un point commun entre les différentes définitions des communautés : on souhaite avoir le maximum d'arêtes à l'intérieur des communautés et donc le minimum à l'extérieur. Nous proposons de contribuer à ce domaine en étudiant un problème de partition de sous-graphe dense, où la densité représente la cohésion de chaque partie, du point de vue de la complexité informatique.

Ce document est organisé en trois chapitres. Le premier présente le contexte nécessaire à la compréhension du matériel présenté dans cette thèse. Nous présentons d'abord des notions centrales liées à l'anonymisation, aux modèles d'anonymisation, à la vie privée, à l'utilité des données, aux attaques puis des notions sur la partition des communautés et enfin nous définissons les notations liées à la théorie des graphes utilisées dans ce document. Le deuxième chapitre se concentre sur

un modèle classique d'anonymisation des graphes. Nous étudions l'existence d'une solution ainsi que la difficulté NP avec une modification choisie du graphe. Nous concevons également un algorithme d'approximation ainsi qu'un algorithme exact simple. Dans un troisième chapitre, nous étudions un problème de partitionnement en graphes denses. Nous classons la difficulté du problème pour plusieurs classes de graphes, dont une relativement similaire aux graphes de réseaux sociaux (épars). Nous concevons ensuite un schéma d'approximation efficace pour les graphes denses.

Preliminaires

Une base de données graphique est une base de données stockée dans une structure de graphe. Les entités sont généralement représentées par des sommets et leurs attributs par l'étiquette du sommet. De plus, on peut ajouter un lien entre deux entités, représenté par une arête. Ce lien peut également avoir des attributs, stockés dans l'étiquette de l'arête. Ces bases de données sont également appelées bases de données de réseaux sociaux, car la plupart des réseaux sociaux sont stockés dans une base de données de graphes.

En 2008, Liu et Terzi ont proposé une adaptation de l'anonymisation classique de k sur le réseau relationnel pour le contexte de la base de données de graphes, ils l'ont nommée *k-degré-anonymisation.* [44]. Ce modèle est conçu pour le cadre des bases de données de graphes, en particulier pour les graphes de réseaux sociaux. Leur idée est d'identifier les violations possibles dans ce contexte et de proposer une solution dédiée. Ils ont regroupé les violations de la vie privée dans les réseaux sociaux en trois catégories : 1) *divulgation de l'identité* : l'identité de l'individu qui est associé au nœud est révélée, il s'agit de la même définition que pour les bases de données relationnelles ; 2) *divulgation des liens* : une relation sensible entre deux individus est révélée (l'adversaire connaît l'existence d'un bord) ; et 3) *divulgation du contenu* : la confidentialité des données associées à chaque nœud est violée (c'est-à-dire le contenu de l'étiquette). Si les lignes de la base de données ne sont pas indépendantes, c'est-à-dire s'il existe un attribut qui prend pour valeur une autre ligne, la divulgation d'un lien dans une base de données graphique peut être considérée comme la divulgation de ce type d'attribut dans une base de données relationnelle. Si cet attribut n'existe pas, la divulgation du lien n'a aucun sens dans les bases de données relationnelles. La divulgation de contenu peut être associée à la divulgation d'attributs. L'anonymat à k-degré se concentre sur la divulgation de l'identité, ce qui n'implique rien pour les autres divulgations (comme nous le voyons dans l'attaque d'homogénéité par exemple). Pour gérer la divulgation du contenu, ils se sont référés à des méthodes d'anonymisation connues, à des techniques standard d'exploration de données préservant la vie privée [1]. Pour se prémunir contre la divulgation des liens, ils se sont référés aux techniques de la communauté d'exploration de liens [29, 61].

Communauté de réseau social

Les réseaux représentant les liens sociaux qui unissent les individus sont étudiés depuis longtemps par les sociologues et les informaticiens [53, 59]. L'émergence des

nouvelles technologies de télécommunication a offert de nouveaux modes d'interaction entre les individus ainsi qu'une facilité d'observations externes pour les chercheurs. Les exemples les plus connus sont le réseau téléphonique et l'Internet. Dans ces réseaux sociaux, il est facile d'observer les interactions de millions d'individus. Dans ce contexte, les communautés peuvent prendre la forme d'un cercle d'amis ou de famille, de personnes partageant des intérêts communs, etc.

En 2008, Blondel et al. [10] ont analysé le réseau formé par les interactions des différents utilisateurs d'un opérateur téléphonique belge. Le graphe contient 2,6 millions de sommets et les arêtes sont pondérées par le temps de communication cumulé entre les deux sommets adjacents (voir Figure 3.8). Comme nous l'avons dit précédemment, la définition précise de la communauté reste un parti pris des auteurs, mais elle partage certaines propriétés communes. En effet, ils ont réussi à extraire 261 groupes de plus de cent sommets qui sont séparés en deux groupes spécifiques : les francophones et les néerlandophones. Ceci est bien sûr totalement improbable dans un graphe uniformément dessiné. Nous pouvons voir qu'un réseau social qui reflète la connexion de la société humaine est loin d'être un graphe aléatoire.

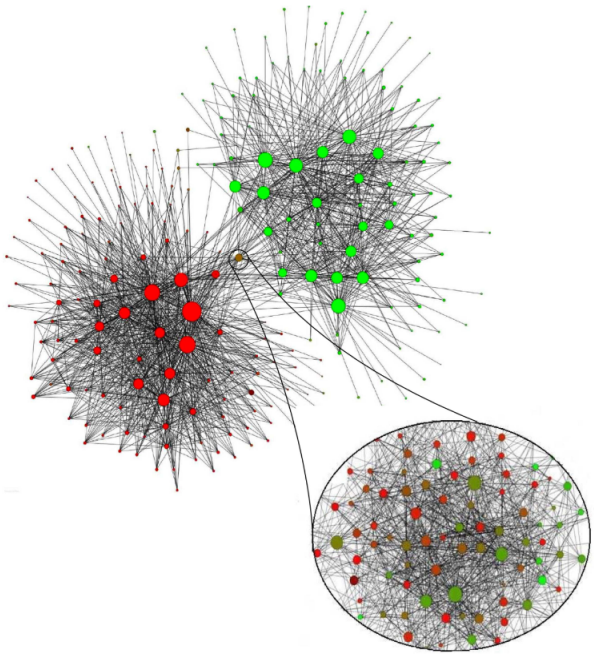


Figure 3.8: Télécommunications belges : Deux langues, deux mondes et très peu de connexions.

Red et al. [51] ont utilisé des données Facebook pour reconstruire un réseau relationnel entre des étudiants de différentes universités américaines (voir Figure 3.9). L'objectif était d'étudier le lien entre la vie en ligne et hors ligne. Ils ont constaté que les communautés étaient structurées autour des classes de terminale ou des dortoirs selon l'université (et la présence de dortoirs).

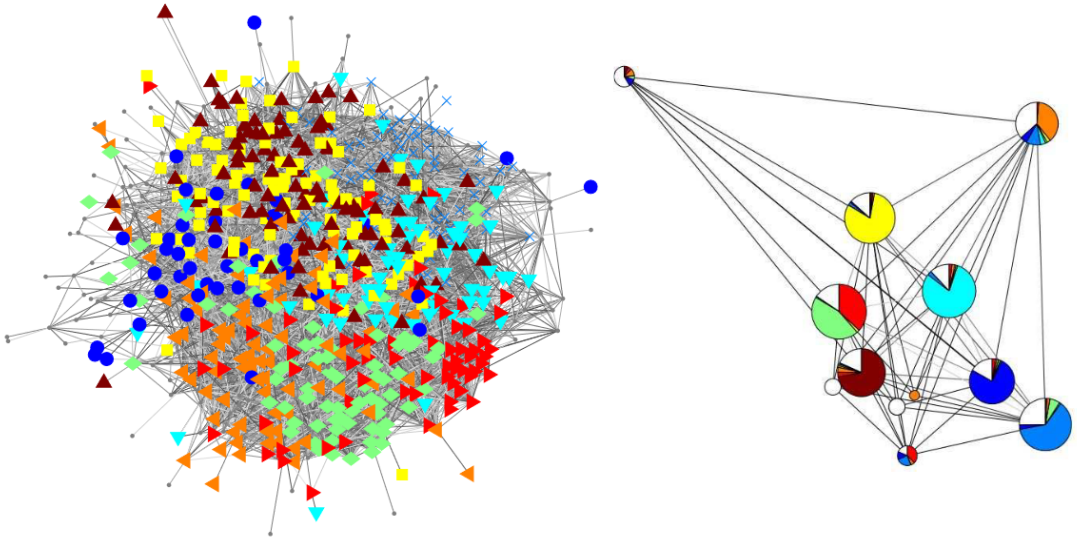


Figure 3.9: Réseau d'amis facebook des étudiants de Caltech, les couleurs/formes indiquent les dortoirs (à gauche). Visualisation des communautés par rapport à l'affiliation des dortoirs (à droite). Les diagrammes circulaires représentent la proportion du nombre de liens vers chaque communauté, y compris elle-même.

Une première définition de la communauté : la clique et ses relaxations

Une première définition, quelque peu restrictive, de Luce et Perry [47] stipule qu'un membre d'une même communauté est apparenté à tous les autres. Cependant, on constate que cette définition est extrêmement restrictive : un sous-groupe de mille personnes interconnectées à quelques exceptions près semble être un groupe cohésif mais n'est pas une communauté avec cette définition. De plus, il a été constaté que les communautés ne sont pas des groupes symétriques d'individus mais une hiérarchie dans le sens où il existe des sommets centraux fortement connectés et des sommets périphériques faiblement connectés au sein d'une même communauté [53, 59]. Enfin, le problème de trouver une clique de taille maximale appartient aux 21 problèmes originaux de Richard Karp, ainsi que le partitionnement d'un graphe avec le nombre minimal de cliques : [39] (voir Figure 3.10). Certains travaux ont été proposés en se basant sur une relation de la notion de clique, appelée t-clique par le sociologue [46, 2] (c'est-à-dire sous-ensemble de sommets de diamètre maximal t dans G). La notion de k-club (i.e. sous-graphe de diamètre maximum k) peut surmonter cette faiblesse mais le problème est NP-hard [34].

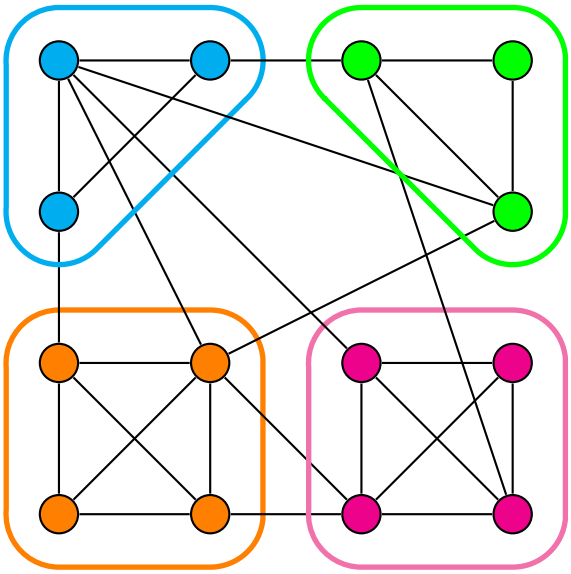


Figure 3.10: Partition into clique

Formalisation de la notion de communauté : fonction de qualité

Dans cette partie, la communauté est définie autour de la notion "il y a plus d'arêtes à l'intérieur qu'à l'extérieur". Cependant cette définition est assez ouverte et on peut trouver beaucoup de définitions qui satisfont ce postulat. D'autre part le graphe entier satisfait systématiquement cette propriété, il faut donc trouver un moyen d'optimiser la taille des communautés trouvées. La solution la plus étudiée consiste à considérer les communautés comme la sortie d'un algorithme donné, sans définition préalable. La plupart de ces algorithmes reposent sur la notion de degré interne et externe. Considérons P une partition des sommets d'un graphe G. Pour tout sommet v d'une partie C de P, nous définissons $v_{int} = |\{u \mid u \in C \wedge u \in \mathcal{N}(v)\}|$ et $v_{ext} = |\{u \mid u \notin C \wedge u \in \mathcal{N}(v)\}|$. De la même manière, nous définissons $C_{int} = \frac{\sum_{v \in C} v_{int}}{|C| \times (|C|-1)}$ et $C_{ext} = \frac{\sum_{v \in C} v_{ext}}{|C| \times (n-|C|)}$ (voir Figure 3.11).

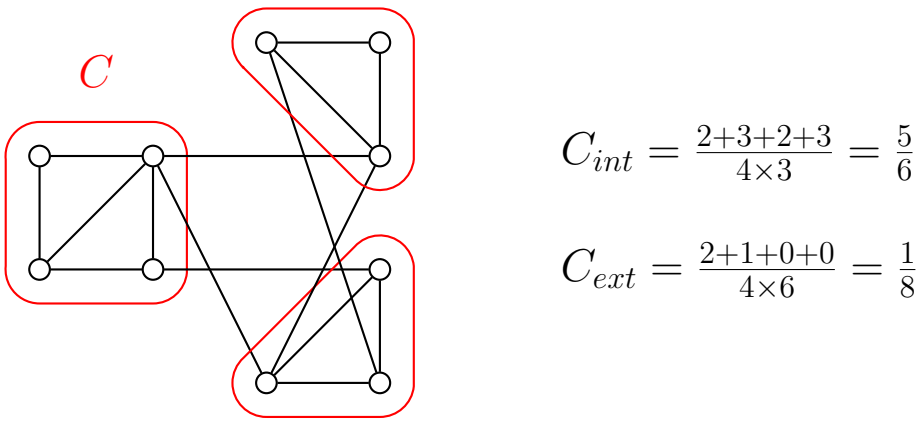


Figure 3.11: Connexions internes et externes d'une communauté C.

Intuitivement, une bonne communauté serait un sous-ensemble C de sommets où C_{int} est beaucoup plus grand que C_{ext}. Trouver un bon compromis entre ces deux quantités est le but de la plupart des algorithmes de détection de communauté. Par exemple, Mancoridis et al. ont essayé de trouver une partition qui maximise la somme des différences $C_{int} - C_{ext}$ de chaque partie dans [49].

D'autre part, une autre condition naturelle nécessaire est la connectivité des communautés. Elle est souvent induite par l'optimisation des compromis mais parfois elle doit être imposée comme une contrainte. En utilisant cette condition de base, de nombreuses définitions de communautés ont été proposées en fonction de l'application.

Partitionnement en communautés à l'aide de la fonction de qualité.

Historiquement, les chercheurs se sont principalement concentrés sur le découpage d'un graphe en un nombre prédéfini de clusters (conception semi-conductrice [40], traitement parallèle [50] et apprentissage non supervisé [45] (k-means)). Dans le contexte des communautés, nous ne connaissons généralement pas le nombre de communautés à l'avance et le fait de l'imposer pourrait donner des résultats indésirables, comme la fusion ou la scission de groupes cohésifs.

Nous sommes donc intéressés à fournir des algorithmes qui divisent le graphe en un nombre indéfini de parties. Bien qu'il soit facile de diviser un graphe en communautés cohésives, toutes les partitions ne sont pas intéressantes. Afin de comparer différentes partitions, nous pouvons utiliser un critère qui les classe entre elles. Souvent, ce critère est matérialisé par ce que l'on appelle une fonction de qualité, c'est-à-dire une fonction qui prend en entrée une partition du graphe et renvoie une valeur dans ce contexte. Fondamentalement, les partitions avec un score élevé sont meilleures que les partitions avec un score faible, donc nous allons naturellement chercher une partition qui le maximise. Cependant, il est important de garder à l'esprit que le choix du critère influence grandement la forme des partitions trouvées et sera donc lié à l'application souhaitée.

Rotations d'arête pour l'anonymisation des degrés

Dans ce chapitre, nous étudions la complexité informatique de Min Anonymous-Edge-Rotation défini comme suit : une entrée du problème est un graphe non orienté $G = (V, E)$ avec n sommets et m arêtes et un entier $k \leq n$. Le but est de trouver la plus courte séquence de rotations d'arêtes qui transforme G en un graphe *k-degré-anonyme*, si une telle séquence existe. Un graphe anonyme à k-degrés est un graphe tel que pour chaque degré possible, il y a soit 0 sommets de ce degré, soit au moins k. Dans un premier temps, nous allons définir les notions utilisées dans ce chapitre et faire le point sur l'état de l'art de l'anonymisation à k-degrés.

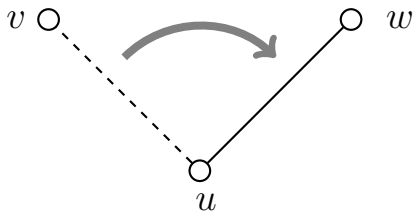


Figure 3.12: An edge rotation (uv, uw) from uv to uw

Dans ce chapitre, nous supposons que tous les graphes sont non orientés, sans boucles ni arêtes multiples, et pas nécessairement des graphes connectés. Soit $\mathbf{G}(n, m)$ l'ensemble de tous les graphes avec n sommets et m arêtes.

Definition 3.5.1. *Soit G, G' dans $\boldsymbol{G}(n, m)$. On dit que G' peut être obtenu à partir de G par une **rotation d'arête** (uv, uw) si $V(G) = V(G')$ et s'il existe trois sommets distincts u, v et w dans G tels que $uv \in E(G)$, $uw \notin E(G)$, et $E(G') = (E(G) \setminus \{uv\}) \cup \{uw\}$, voir Figure 2.1.*

Remark 3.5.1. Soit G un graphe. Pour les sommets u, v, w dans G la rotation des arêtes (uv, uw) modifie G en un graphe G' tel que $d_{G'}(v) = d_G(v) - 1$, $d_{G'}(w) = d_G(w) + 1$, et le degré des autres sommets n'est pas modifié. Définissons une modification du degré $(+1, -1)$ de la séquence de degrés $D = (d_1, \ldots, d_n)$ de telle sorte que $d_i := d_i + 1$, $d_j := d_j - 1$ pour deux indices quelconques i, j tels que $i, j \in \{1, \ldots, n\}$. Notez que chaque rotation d'arête correspond à une modification de $(+1, -1)$ degré, mais pas l'inverse.

Definition 3.5.2. *Une suite d'entiers $D = (d_1, d_2, \ldots, d_n)$ est appelée k-**anonymous** où $k \in \{1, \ldots, n\}$, si pour chaque élément d_i de D il existe au moins $k - 1$ autres éléments dans D avec la même valeur. Un graphe G est appelé k-**degré-anonyme** si sa suite de degrés est k-**anonyme**. Les sommets de même degré correspondent à une **classe de degrés**.*

Dans ce chapitre, nous étudions le problème d'anonymisation suivant :

Min-Rotation d'arêtes anonymes
Input: (G, k) où $G = (V, E)$ est un graphe non orienté et k un entier positif, $k\{1, \ldots, \|V\|\}$. et k un entier positif, $k \in \{1, \ldots, \|V\|\}$.
Output: S'il existe une solution, trouver une séquence d'un nombre minimum $\ell + 1$ de graphes $G_0 = G, G_1, G_2, \ldots, G_\ell$ tels que G_{i+1} peut être obtenu à partir de G_i par une rotation d'arête, et G_ℓ est k-degré-anonyme.

Notez qu'une solution au problème Min Anonymous-Edge-Rotation peut ne pas exister pour toutes les instances. Par exemple, si G est un graphe complet sans arête, $K_n \setminus \{e\}$, $n \geq 6$, alors il n'y a pas de solution pour un tel graphe G et $k = 3$. Par conséquent, nous nous intéressons uniquement à l'étude des **instances faisables** (G, k) définies comme une instance pour laquelle il existe une solution à

Les résultats de cette section ont été publiés dans [6, 7].

Min Anonymous-Edge-Rotation. Notre étude initiale des conditions suffisantes de faisabilité est présentée dans la section 3.5.

Évidemment, puisque tous les graphes sont anonymes à 1 degré, nous ne sommes intéressés que par les cas où $k \geq 2$.

La version de décision associée à Min Anonymous-Edge-Rotation est définie comme suit pour une instance faisable (G, k) :

Anonymous-Edge-Rotation
Input: (G, k, r) où $G = (V, E)$ est un graphe non orienté, $k \in \{1, \ldots, |V|\}$, et r un entier positif.
Question: Existe-t-il une séquence de graphes $\ell + 1$ $G_0 = G, G_1, G_2, \ldots, G_\ell$ telle que $\ell \leq r$, G_{i+1} peut être obtenu à partir de G_i par une rotation d'arête, et G_ℓ est k-degré-anonyme ?

Nous considérons également le problème Min Anonymous-Edge-Rotation dans les classes de graphes restreintes, par exemple les arbres. Dans ce cas, nous exigeons que tous les graphes de la séquence $G_0, \ldots, G_\ell$ soient de la même classe de graphes. Notez que le problème peut également être étudié sans cette condition, mais les résultats peuvent être différents.

Nous allons maintenant résumer les différents résultats sur l'anonymisation à k-degrés. La notion a d'abord été étudiée dans la littérature à travers des opérateurs plus classiques comme l'ajout ou la suppression d'arêtes. Nous nous concentrerons ensuite sur les travaux réalisés avec la rotation d'arêtes.

Ce chapitre est organisé comme suit. L'étude de la faisabilité est initiée dans Section 3.5. Le Section 3.5 présente la preuve de la dureté NP de Anonymous-Edge-Rotation. Dans la Section 3.5, nous étudions les propriétés des séquences de degrés anonymes spécifiques à k-degrés qui sont utilisées dans la Section 3.5 pour présenter un algorithme d'approximation à 2 en temps polynomial et dans la Section 3.5 pour établir un algorithme en temps polynomial pour les arbres. De plus, dans Section 3.5 nous considérons le cas $k = n$ dans les graphes généraux. Quelques conclusions sont données à la fin du chapitre.

Faisabilité

Le problème Min Anonymous-Edge-Rotation n'a pas de solution pour chaque instance d'entrée. Il n'est pas difficile de voir que si un graphe est "presque" complet ou "presque" vide, alors il n'y a que des options restreintes sur le nombre de classes de degrés différents et donc une solution peut ne pas exister.

Nous montrons d'abord que nous pouvons atteindre n'importe quel graphe à partir de n'importe quel autre graphe avec le même nombre de sommets et d'arêtes via des rotations d'arêtes. Ensuite, nous présentons quelques conditions suffisantes pour qu'une instance soit réalisable. Enfin, nous montrons qu'une solution du problème existe pour tous les $k \leq \frac{n}{4}$, où n est l'ordre du graphe.

Le théorème suivant montre des propriétés importantes sur les rotations des arêtes. Le résultat a déjà été prouvé dans [16], mais en raison de la simplicité de notre approche, nous présentons ici une autre preuve.

Theorem 3.5.1. *Pour deux graphes quelconques $G, G' \in \boldsymbol{G}(n,m)$, nous pouvons transformer G en G' en utilisant une séquence de rotations d'arêtes.*

Corollary 3.5.1. *Pour deux graphes quelconques G, G' in $\boldsymbol{G}(n,m)$, la distance entre les arêtes de G et G' est bornée par $2m$.*

Theorem 3.5.2. *Soit G dans $\boldsymbol{G}(n,m)$ tel que $\frac{n}{2} \leq m \leq \frac{n(n-3)}{2}$ et $n \geq 4$. Alors il existe une solution réalisable pour le problème* Min Anonymous-Edge-Rotation*, donc un graphe anonyme de degré k $G' \in \boldsymbol{G}(n,m)$, pour tout $k \leq \frac{n}{4}$.*

Nous étendons maintenant l'étude de faisabilité au cas $k = n$ pour lequel nous obtenons des conditions nécessaires et suffisantes.

Theorem 3.5.3. *Soit $G \in G(n,m)$ pour certains entiers positifs n et m. Alors (G,n) est une instance réalisable de* Min Anonymous-Edge-Rotation *si et seulement si $\frac{2m}{n}$ est un entier.*

NP-difficulté

Dans cette section, nous montrons que la version décisionnelle de Min Anonymous-Edge-Rotation, le problème Anonymous-Edge-Rotation, est NP-dur. La preuve est basée sur une réduction de la version restreinte d'un problème d'ensemble de couverture, Exact Cover By 3-Sets, qui est connu pour être NP-complet [27].

Theorem 3.5.4. Anonymous-Edge-Rotation *est NP-hard même dans le cas $k = \frac{n}{q}$ où n est l'ordre du graphe G pour une instance d'entrée (G,k,r) et q est un nombre fixe supérieur ou égal à 3.*

Borne inférieure pour les rotations

Dans cette section, nous supposons que (G,k) est une instance réalisable. Pour toute instance de ce type, nous définissons une séquence de degrés anonymes S_{bound} qui peut être calculée en temps polynomial si $k = \theta(n)$. Nous montrons qu'avec les modifications de degré $(+1,-1)$ (Remark 3.5.1) le graphe G peut être transformé en un graphe anonyme de degré k G' avec la séquence de degré S_{bound} en utilisant au plus le double de rotations d'arêtes comme dans une solution optimale de Min Anonymous-Edge-Rotation pour (G,k).

Notez qu'en général, une modification de degré $(+1,-1)$ ne correspond pas à une rotation d'arête, mais comme nous le montrerons plus tard dans la section sec:trees, c'est vrai pour les arbres.

Theorem 3.5.5. *Soit (G,k) une instance réalisable pour le problème* Min Anonymous-Edge-Rotation*. Soit OPT une solution optimale qui est un ensemble minimal de rotations qui transforment G en un graphe anonyme de k degrés G'. Alors $\sum_{i=1}^{n} |S_G[i] - S_{bound}[i]| \leq 2|OPT|$.*

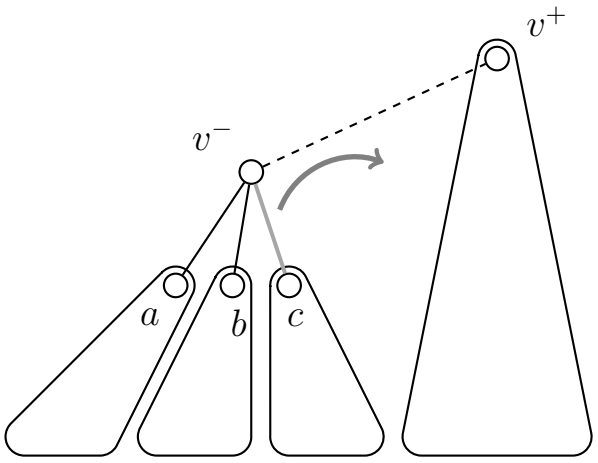


Figure 3.13: Transformation de T vers T'

Approximation

Dans cette section, nous montrons que sous certaines contraintes sur le nombre d'arêtes et k, il existe un algorithme d'approximation en temps polynomial 2-approximation pour le problème Min Anonymous-Edge-Rotation pour toutes les entrées réalisables (G, k).

Theorem 3.5.6. *Le problème* Min Anonymous-Edge-Rotation *est approchable en temps polynomial à 2 pour toutes les instances (G, k), où $k \leq \frac{n}{4}$ où $k = \theta(n)$ et G est le graphe avec n sommets et m arêtes, où $\max\{\frac{n}{2}, (1+\frac{n}{4k}+\frac{n}{k\Delta})^2\Delta^2\} \leq m \leq \frac{n(n-3)}{2}$, et la constante c est définie comme $c = \lfloor \frac{n}{k} \rfloor$.*

Cas polynomiaux

Comme le montre la section 3.5, le problème Min Anonymous-Edge-Rotation est NP-hard même pour $k = \frac{n}{q}$ et $q \geq 3$ est une constante fixe où n est l'ordre d'un graphe d'entrée. Dans cette section, nous montrons que le problème peut être résolu en temps polynomial sur les arbres lorsque $k = \theta(n)$ ou dans le cas d'un graphe quelconque lorsque $k = n$.

Arbres

Pour un arbre $T = (V, E)$ ayant pour racine un sommet r, pour tout $v \in V$, $v \neq r$, $child(v)$ est un sommet qui est un voisin de v ne se trouvant pas sur le chemin allant de r à v.

Lemma 3.5.1. *Soit $T = (V, E)$ un arbre et v^-, v^+ sommets de V tels que v^- n'est pas une feuille et v^+ n'est pas un sommet universel. universel. Ensuite, en utilisant une rotation, nous pouvons transformer T en un arbre T' tel que $d_{T'}(v^-) = d_T(v^-) - 1$ et $d_{T'}(v^+) = d_T(v^+) + 1$.*

Theorem 3.5.7. *Le problème* Min Anonymous-Edge-Rotation *est soluble en temps polynomial pour toute instance (T, k) où T est un arbre d'ordre n, $k \leq \frac{n}{4}$ et tel que $c = \lfloor \frac{n}{k} \rfloor$ est une constante, donc $k = \theta(n)$.*

Une classe de degré, $k = n$.

Dans cette partie, nous montrons que Min Anonymous-Edge-Rotation est soluble en temps polynomial pour les instances où k coïncide avec le nombre de sommets du graphe, c'est-à-dire que tous les sommets doivent être dans la même classe de degré.

Lemma 3.5.2. *Soit $G = (V, E)$ un graphe et $u, v \in V$. Si $\mathcal{N}_G(u) \nsubseteq \mathcal{N}_G(v)$, alors il existe une rotation des arêtes qui conduit à un graphe G' tel que $d_{G'}(u) = d_G(u) - 1$ et $d_{G'}(v) = d_G(v) + 1$.*

Remark 3.5.2. *Soit $G = (V, E)$ un graphe, pourtoutu, $v \in V$, si $d_G(u) > d_G(v)$, alors il existe une rotation des arêtes qui conduit à un graphe G' tel que $d_{G'}(u) = d_G(u) - 1$ et $d_{G'}(v) = d_G(v) + 1$.*

Lemma 3.5.3. *Soit (G, n) une instance de* Min Anonymous-Edge-Rotation *; où $G \in \boldsymbol{G}(n, m)$ pour certains entiers positifs m, n, et $\frac{2m}{n}$ est un entier. Alors la valeur optimale de Min Anonymous-Edge-Rotation ; sur (G, n) est $\frac{\sum_{w \in V} |d_G(w) - 2m/n|}{2}$.*

Theorem 3.5.8. *Le problème* Min Anonymous-Edge-Rotation *est soluble en temps polynomial pour les instances (G, k) lorsque $k = n$, où n est l'ordre du graphe G.*

Conclusion

Nous commençons l'étude de la complexité du problème Min Anonymous-Edge-Rotation dans lequel la tâche est de transformer un graphe donné en un graphe anonyme de degré k en utilisant un nombre minimum de rotations d'arêtes. Comme nous avons pu prouver la dureté NP dans le cas où le nombre de sommets k dans chaque classe de degré est $\theta(n)$, des recherches supplémentaires pourraient explorer des résultats de dureté plus forts ou des cas où k est une constante. Une prochaine étape de recherche pourrait inclure la relaxation de la condition sur le nombre d'arêtes dans l'algorithme de 2-approximation présenté ainsi que l'extension des classes de graphes dans lesquelles le problème Min Anonymous-Edge-Rotation peut être résolu en temps polynomial. Comme le problème n'a pas de solution pour tous les graphes et toutes les valeurs possibles de k, notre étude initiale de faisabilité couvre une grande partie des instances. Les extensions des résultats sont encore possibles, dans le sens des conditions nécessaires et suffisantes.
Une autre voie intéressante pour continuer pourrait être d'étudier le même modèle d'anonymisation mais avec l'opérateur de suppression/addition d'arête qui consiste à supprimer une arête pour la réinjecter à un autre endroit où une arête était manquante. Il présente de nombreuses similitudes avec la rotation d'arêtes et leurs distances d'édition sont liées. Les résultats positifs semblent faciles à étendre, cependant la preuve de dureté est trop spécifique à la rotation, une nouvelle preuve pourrait être conçue.

Communities and Dense Graph Partitioning

Dans ce chapitre, nous étudions le problème Max Dense Graph Partition consistant à trouver une partition $\mathcal{P} = \{V_1, \ldots, V_k\}$, $k \geq 1$, d'un graphe non orienté donné G, telle que la densité de la partition, notée $d(\mathcal{P})$, soit maximisée. Nous considérons une définition classique de la densité d'un sous-graphe induit par un sous-ensemble S de sommets (voir, par exemple, [21, 31]) donnée par le rapport entre le nombre d'arêtes et le nombre de sommets dans S. La densité d'une partition est la somme des densités de toutes ses parties. En effet, lorsque le nombre de classes est donné, le problème est une généralisation d'une partition en k cliques. Nous abordons donc le problème Max Dense Graph Partition de trouver une partition de densité maximale, sans fixer le nombre de classes de la partition.

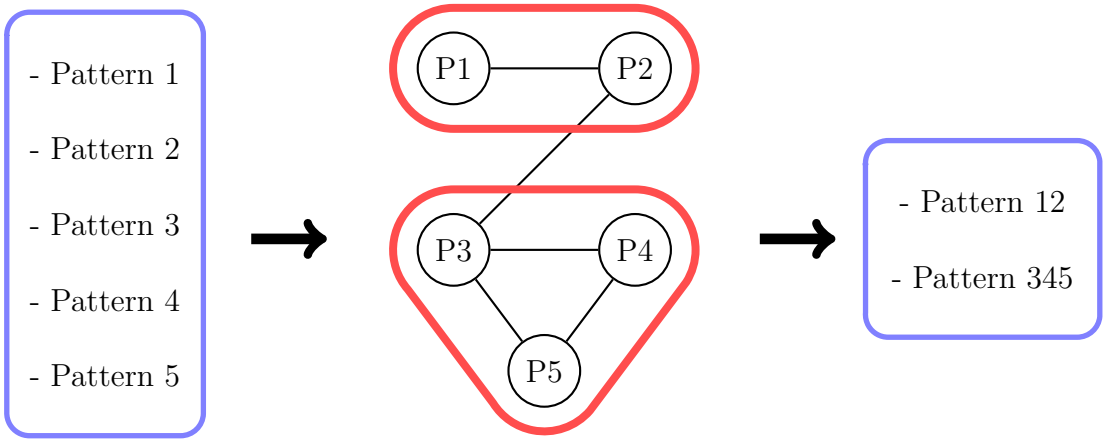
Preliminaires

Pour cette définition de la densité, il existe plusieurs articles sur la recherche du sous-graphe le plus dense. Ce problème a été démontré soluble en temps polynomial par Goldberg [31] mais si la taille du sous-graphe est une partie de l'entrée, le problème appelé k-Sous-graphe le plus dense devient NP-hard même restreint aux graphes bipartites ou chordaux [18]. L'approximabilité de k-Sous-graphe le plus dense a également été étudiée, voir [41, 25, 9].

Motifs médicaux

Le problème a été introduit par Darlay et al. [21] dans le contexte de l'analyse de données médicales. Durant son doctorat [20], ils ont été confrontés à un problème de détection de communautés dans des données médicales pour la recherche clinique. Ils ont utilisé la méthode d'analyse logique des données [11] où un grand ensemble de motifs est généré (voir Figure 3.14a), un motif étant les caractéristiques des patients ayant des propriétés similaires pour la pathologie étudiée. Leur objectif est d'identifier les patrons similaires pour les fusionner afin d'en réduire le nombre. L'idée est de partitionner le graphe en communautés de motifs similaires (voir Figure 3.14b). Les motifs appartenant à la même communauté seront ensuite assimilés (voir Figure 3.14c).

Les résultats de cette section ont été publiés dans [5] (version arXiv).



(a) Génération d'un grand nombre de motifs

(b) Graphe de similarité et sa partition en sous-graphes denses

(c) Ensemble de motifs compressés

Figure 3.14: Processus de compression des motifs

Darlay et al. ont étudié Max Dense Graph Partition, et son complément Min Sparse Graph Partition. Ils ont défini la sparsité d'une partition $\mathcal{P}$ comme $F(\mathcal{P}) = \frac{|\mathcal{P}|}{2} + d(\mathcal{P})$ et le problème Min Sparse Graph Partition comme étant la recherche d'une partition d'un graphe non orienté donné G telle que la sparsité de la partition soit minimisée. Observez que ces deux problèmes Max Dense Graph Partition et Min Sparse Graph Partition sont des duaux dans le sens où résoudre le premier sur un graphe G est identique à résoudre le second sur le complément de G. Dans [21], il est montré que les deux problèmes sont NP-complets, et qu'il n'existe pas d'approximation à facteur constant pour Min Sparse Graph Partition à moins que $P = NP$. De plus, un algorithme en temps polynomial est donné pour Max Dense Graph Partition sur les arbres.

Nous soulignons que leur preuve de NP-complétude est une réduction en temps polynomial de k-Coloring. Par construction, la même réduction en partant de 3-Coloring sur les graphes de degré au plus 4 (prouvé NP-complet dans [28]) donne comme instance de Max Dense Graph Partition un graphe sur n sommets et de degré minimum plus grand que $n - 4n^{4/5}$. Il s'ensuit que Max Dense Graph Partition est NP-complet restreint aux graphes de degré minimum $n - 4n^{4/5}$.

Coalitions et communautés : Jeu hédonique

Le jeu hédonique est un cadre naturel pour étudier l'aspect formel de la formation de coalitions. Une coalition peut être vue comme une communauté où le lien entre les individus est un intérêt commun. Ce jeu modélise la formation de coalitions lorsque les joueurs ont une préférence pour le groupe auquel ils appartiennent. Fondamentalement, on veut former des coalitions de personnes très proches, ce qui est proche de l'approche de partitionnement en sous-graphes denses. Dans un problème d'optimisation lié au jeu hédonique simple, symétrique et fractionnaire, Aziz et al. s'intéressent à la recherche d'une partition qui maximise le bien-être utilitaire [3], ce qui est équivalent au problème Dense Graph Partition. Dans cette partie,

nous allons d'abord montrer que les problèmes sont équivalents, puis présenter leurs résultats.

La définition de base d'un jeu hédonique est un ensemble N de n éléments et un ensemble de relations complètes et transitives $\succsim = (\succsim_1, \cdots, \succsim_n)$ qui modélisent les préférences de chaque joueur pour les différentes coalitions (i.e. sous-ensemble de N).

Le jeu est dit *fractionnel* si pour chaque joueur i on définit une fonction d'utilité $v_i : N \to \mathbb{R}$ qui associe une valeur aux autres joueurs. Les coalitions sont considérées comme un ensemble d'agents distincts. Pour la convertir en un jeu hédonique, une fonction intermédiaire est nécessaire : laissez $v_i(S) = \frac{\sum_{j \in S} v_i(j)}{|S|}$ être la participation de l'agent i dans la coalition S. En utilisant cette fonction, un jeu hédonique est dit *fractionnel* si pour chaque joueur $i \in N$, pour toutes les coalitions $S, T \subseteq N$, $S \succsim_i T$ si et seulement si $v_i(S) \geq v_i(T)$.

Lemma 3.5.4. *Un jeu hédonique simple, symétrique et fractionnaire $(N, \succsim)$ peut être représenté par un graphe $G = (N, \{ij \subseteq N \mid v_i(j) = 1\})$.*

Le *bien-être utilitaire* d'une partition est la somme de l'utilité de ses agents. Plus formellement, cela signifie $\sum_{P \in \mathcal{P}} \sum_{i \in P} v_i(P)$.

Lemma 3.5.5. *Le bien-être utilitaire d'une partition $\mathcal{P}$ de N est égal à deux fois la densité de la même partition $\mathcal{P}$ dans G.*

Nous déduisons qu'une partition qui maximise le bien-être utilitaire dans un jeu hédonique fractionnaire symétrique simple $(N, \succsim)$ maximise également la densité d'un graphe $G = (N, \{ij \subseteq N \mid v_i(j) = 1\})$ et inversement.

Dans [3] Aziz et al. ont prouvé que le problème est NP-complet même pour les graphes 3-parties. D'autre part, ils ont montré qu'un appariement maximal fournissait une approximation de 2 dans le cas général. À notre connaissance, il n'existe aucun autre travail sur cette version spécifique du jeu hédonique.
Nos contributions.

La vue d'ensemble suivante résume les résultats obtenus dans ce chapitre concernant la Partition de graphique à densité maximale.

- Max Dense Graph Partition est trivialement soluble sur les graphes de degré maximum 2, nous prouvons son caractère NP-hardness pour les graphes 3-réguliers (cubiques).

- Nous établissons que sur les graphes complets bipartis, une partition optimale consiste en une partie, c'est-à-dire le graphe entier. De plus, si la taille des deux ensembles indépendants est un nombre relativement premier, alors cette solution optimale est unique. Nous utilisons ce résultat pour montrer que Max Dense Graph Partition est $W[2]$ difficile en ce qui concerne (une borne supérieure sur) le nombre de clusters dans une solution optimale sur les graphes bipartis denses. Notre réduction est polynomiale et implique donc en particulier la NP-dureté de Max Dense Graph Partition sur les graphes bipartis denses.

Max Dense Graph Partition est triviale sur les graphes complets puisque la solution optimale est le graphe entier comme une partie de la partition. De plus, comme nous l'avons expliqué précédemment, il est NP-hard sur les graphes de degré minimum $n - 4n^{4/5}$. Nous montrons que pour les graphes de degré minimum $\geq n-3$, le problème est soluble en temps polynomial et toute solution optimale a deux parties. De plus, sur les graphes $(n-4)$-réguliers, le problème devient NP-hard.

- Nous montrons que Max Dense Graph Partition admet une $(1+\varepsilon)$-approximation pour tout $\varepsilon > 0$ sur les graphes $(n-4)$-réguliers, améliorant la 2-approximation sur les graphes généraux [3].

Ce chapitre est organisé comme suit. Les notations et les définitions formelles sont données dans la sous-section ci-dessous. L'étude des graphes bipartis (denses) est établie dans la section 3.5. La section 3.5 présente les résultats sur les graphes cubiques. Dans la section 3.5, nous étudions les graphes denses. les graphes denses. Quelques conclusions sont données à la fin du chapitre.

Graphes bipartis denses

Dans cette section, nous montrons que Max Dense Graph Partition a une solution triviale sur les graphes bipartis complets. De plus, en utilisant ce résultat, nous montrons que le problème est NP-dur sur les graphes bipartis denses et même $W[2]$-dur en ce qui concerne le nombre de clusters dans une solution optimale comme paramètre.

Dans la première partie, nous considérons un graphe biparti complet $G_{n,m}$ avec les deux sous-ensembles qui sont des ensembles indépendants de taille n et m et nous prouvons d'abord le résultat suivant.

Lemma 3.5.6. *La densité $d(G_{n,m})$ d'un graphe biparti complet $G_{n,m}$ est supérieure ou égale à la densité $d(\mathcal{P})$ de toute partition $\mathcal{P}$ de $G_{n,m}$.*

Il s'ensuit qu'une solution optimale de tout graphe biparti complet est le graphe entier. À partir des calculs de la preuve précédente, nous pouvons déduire inductivement le résultat suivant.

Corollary 3.5.2. *Pour tout graphe biparti complet $G = (A, B, E)$ avec $|A| = n$ et $|B| = m$, une partition $\mathcal{P} = \{V_1, \ldots, V_k\}$ de $A \cup B$ satisfait $d(\mathcal{P}) = \frac{nm}{n+m}$ si et seulement si $G[V_i] = G_{n_i,m_i}$ avec $n_i \neq 0$ et $m_i \neq 0$ et $\frac{n_i}{m_i} = \frac{n}{m}$ pour tout idans$\{1, \ldots, k\}$.*

Par conséquent, pour tout graphe bipartite complet $G_{n,m}$, si n et m sont relativement premiers, l'unique solution optimale de $G_{n,m}$ est le graphe entier. Sinon, plusieurs solutions optimales existent et sont caractérisées exactement par le Corollaire 3.5.2.

Dans la deuxième partie de cette section, nous étudions le rôle du nombre d'ensembles dans une solution optimale pour Dense Graph Partition. Nous considérons spécifiquement le paramétrage dans le sens suivant. Pour donner formellement un paramètre en entrée, nous considérons des instances de la forme $((G, r), k)$

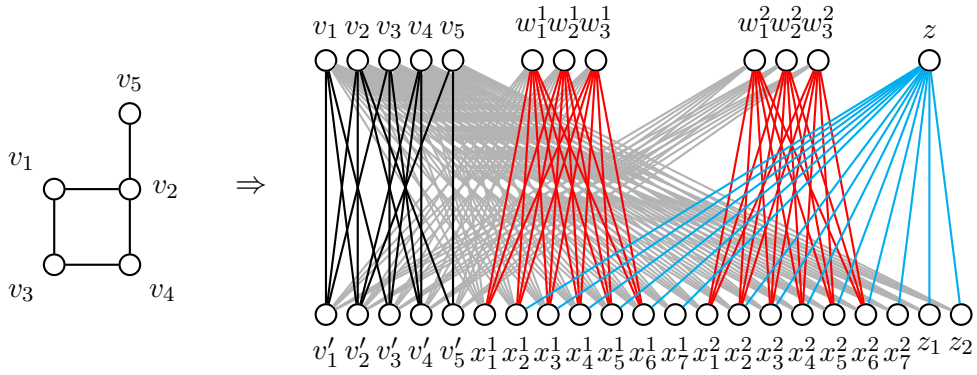


Figure 3.15: Un graphe G, instance de Ensemble dominant et le graphe biparti G' obtenu à partir de G, pour $k = 2$ et $n = 5$.

pour le problème de décision : Existe-t-il une partition $\mathcal{P}$ des sommets de G en au plus k ensembles avec $d(\mathcal{P}) \geq r$? Formellement, l'algorithme peut donc aussi répondre non, si la borne k donnée en entrée n'est pas assez grande pour construire une partition de densité r.

Theorem 3.5.9. *Dense Graph Partition paramétrée par (une borne supérieure sur) le nombre de clusters dans une solution optimale est W[2]-hard sur les graphes bipartis denses.*

Graphes cubiques

Nous montrons que le problème Dense Graph Partition est NP-complet même pour les graphes cubiques en donnant une réduction de Exact Cover By 3-Sets où chaque élément apparaît dans exactement 3 ensembles, noté Restricted Exact Cover By 3-Sets, connu pour être NP-hard par [32].

Restricted Exact Cover By 3-Sets (RX3C)
Input: Un ensemble X d'éléments avec $\|X\| = 3q$ et une collection C de sous-ensembles à 3 éléments de X où chaque élément apparaît dans exactement 3 ensembles ?
Question: C contient-il une couverture exacte pour X, c'est-à-dire une sous-collection $C' \subseteq C$ telle que chaque élément apparaît dans exactement un membre de C' ?

La distinction des cas sur les sous-graphes de $\sigma(I)$ montre :

Lemma 3.5.7. *Pour tout sous-ensemble $S \subseteq V$ des sommets du graphe $\sigma(I)$, les seuls sous-graphes $G[S]$ avec $u(S) \geq \frac{1}{4}$ sont :*

- *un triangle dont tous les sommets sont de type 2 et alors $u(S) = \frac{1}{3}$.*
- *un appariement entre deux sommets de type 2 ou entre deux sommets de types différents et alors $u(S) = \frac{1}{4}$.*

- *le sous-graphe décrit dans la figure 3.4 et ensuite* $u(S) = \frac{1}{4}$.

Remark 3.5.3. *Pour tout sous-ensemble* $S - ensembleqV$ *des sommets du graphe* $\sigma(I)$*, si* v *est de type 2 alors* $u_S(v) \leq \frac{1}{3}$*, sinon* $u_S(v) \leq \frac{1}{4}$.

Avec ces observations sur la construction de $\sigma(I)$, on peut montrer que $I = (X, C)$ est une yes-instance de RX3C si et seulement si $I' = (\sigma(I), d)$ est une yes-instance de Dense Graph Partition qui donne ce qui suit.

Theorem 3.5.10. *Dense Graph Partition est NP-complet sur les graphes cubiques.*

Graphes denses

Dans cette section, nous considérons des graphes $G = (V, E)$ sur n sommets tels que G peut être vu comme $G = K_n - H$ où H est un graphe de petit degré maximum. Les arêtes de H sont appelées *arêtes manquantes* dans G. Nous considérons d'abord des graphes $G = (V, E)$ sur n sommets tels que $\delta(G) \geq n-3$, c'est-à-dire $G = K_n - H$ où H a $\Delta(H) = 2$ et a $q \leq n$ arêtes et montrons que Max Dense Graph Partition est soluble en temps polynomial sur ces graphes.

Theorem 3.5.11. Max Dense Graph Partition *est soluble en temps polynomial sur les graphes* G *avec* n *sommets et* $\delta(G) \geq n - 3$.

Dans la suite de la section, nous considérons des graphes $G = (V, E)$ sur n sommets, $(n-4)$-réguliers, c'est-à-dire $G = K_n - H$ où H est un graphe cubique. Nous montrons que Dense Graph Partition est NP-hard sur les graphes $(n-4)$-réguliers, en montrant une réduction de unCut sur les graphes cubiques, qui est le complément de Max Cut. Ce dernier problème sur les graphes cubiques a été prouvé NP-hard et même non polynomial en temps 1.003-approximable, sauf si P=NP [8].

Min UnCut
Input: Un graphe $G = (V, E)$, un entier k.
Question: Est-ce que G contient une partition de V en deux parties A, B telle que le nombre d'arêtes dont les deux extrémités se trouvent dans la même partie est au plus égal à k ?

Comme nous n'avons pas trouvé de référence pour le résultat suivant dans la littérature.

Lemma 3.5.8. *Soit* $G = (V, E)$ *un graphe cubique. Il existe une partition* $\{A, B\}$ *de* G *avec une coupe de taille au moins égale à* $|V|$ *et elle peut être trouvée en temps polynomial.*

Theorem 3.5.12. Dense Graph Partition *est NP-complet sur* $(n-4)$*-graphes réguliers avec* n *sommets.*

À la fin de cette section, nous montrons qu'une partition en trois cliques fournit une bonne approximation du problème.

Lemma 3.5.9. *Soit $G = (V, E)$ un graphe régulier $(n-4)$ et $\mathcal{P}$ toute partition de V. Alors $d(\mathcal{P}) \leq \frac{n}{2} - 1$.*

Theorem 3.5.13. *Il existe un schéma d'approximation efficace en temps polynomial pour* Max Dense Graph Partition *sur les graphes réguliers $(n-4)$.*

Conclusion

Nous avons poursuivi l'étude de Max Dense Graph Partition initiée par Darlay et al. dans [21] et continuée par Aziz et al. dans [3]. Le problème consiste à identifier les communautés/parties/coalitions qui constituent un graphe à travers le critère de densité. Comme Darlay et al. ont montré que le problème est soluble en temps polynomial sur les arbres, une direction de recherche future intéressante pourrait être d'étudier la complexité du problème paramétrée par l'ensemble d'arêtes de rétroaction minimum. D'autre part, il existe une 2-approximation du problème, nous avons même montré qu'il est possible de faire mieux dans les graphes denses. Pour le moment il n'y a pas de résultats d'inapproximation, des recherches supplémentaires pourraient explorer ce domaine pour donner une meilleure compréhension de la frontière de l'approximation. Toujours dans les graphes denses, il faudrait voir s'il existe un algorithme XP (paramétré par le nombre de parties) ou non afin de terminer le travail commencé sur la W[i]-hardness. Le fait que la fonction objectif densité ne soit pas un entier implique des problèmes d'ordre algorithmique mais aussi pour les preuves. Comme nous l'avons vu dans les preuves de cette section, la simple comparaison de deux quantités peut être une tâche difficile. Il semble donc important de trouver une stratégie algorithmique et de preuve qui puisse surmonter cette difficulté.

Conclusion

Cette thèse présente des résultats sur deux problèmes originaux, un problème d'édition de degrés où la solution est une séquence (ordonnée) d'arêtes liée à l'anonymisation de graphes et un problème de partitionnement optimisant une fonction objectif rationnelle. Nous avons développé des techniques algorithmiques adaptées à ces problèmes et nous avons identifié des cas dans lesquels les problèmes sont devenus solubles en temps polynomial bien qu'ils soient intraitables dans le cas général.

En ce qui concerne le problème de l'anonymisation, deux perspectives ouvertes sont proposées. Premièrement, étendre les résultats sur l'anonymisation à k-degrés sur des modèles plus contraints. Précisément, de nouveaux modèles avancés ont été proposés (comme la i-hop-anonymisation, la k-neighbourhood-anonymisation ou la k-automorphism-anonymisation, voir [14]). Chacune d'entre elles présente un intérêt particulier. Cependant, aucune étude théorique n'a été consacrée à la classification de la complexité du problème de l'anonymisation sur ces modèles. Deuxièmement, ces questions méritent d'être abordées comme une question multidisciplinaire, c'est-à-dire que l'adaptation de méthodes statistiques et probabilistes peut conduire à la résolution de questions intéressantes. Il existe des modèles intermédiaires proposant un compromis entre la précision et la perte de données entre le modèle

d'anonymisation k et le modèle de confidentialité différentielle ϵ (voir [43]). Dans ces modèles, on pourrait diminuer l'epsilon requis de la confidentialité différentielle, ce qui aurait pour effet d'impacter la perte d'utilité des données. Pour compenser l'affaiblissement causé par la diminution du ϵ, on pourrait k-anonymiser, ce qui est généralement moins destructeur pour les données (avec une faible valeur de k). Enfin, établir un lien théorique entre ces deux modèles peut, d'une part, valider les résultats expérimentaux obtenus dans [37], d'autre part, conduire à des garanties équivalentes intéressantes (telles que si un graphe est k-degré-anonyme alors il est $f(k)$-différemment privé) sur certaines classes de graphes restreintes.

En ce qui concerne la densité, de nombreuses variations pourraient être proposées pour répondre à d'autres besoins de détection de communautés. Il serait intéressant de pouvoir exiger une densité minimale des parties (ou une utilité minimale des sommets) pour garantir une qualité minimale pour chaque communauté. Le problème est que nous avons besoin d'une limite adaptative pour résoudre le problème suivant : une limite élevée dans un graphe avec des composantes éparses peut conduire à la non-existence de solutions. D'un autre côté, une borne inférieure sur la taille des parties pourrait être intéressante, avec la formulation actuelle il est tout à fait possible de retourner des communautés d'un seul sommet isolé dans un graphe régulier.

RÉSUMÉ

Dans cette thèse nous étudions la complexité algorithmique de problèmes d'optimisation liés à l'étude des réseaux sociaux. Un réseau social peut être modélisé par un graphe dans lequel les sommets représentent les membres et les arêtes, leurs relations. Nous nous intéressons à deux problématiques différentes, l'anonymisation de graphes et la détection de communautés.
L'anonymisation de graphes a déjà été étudiée et de nombreuses formalisations du concept ont été proposées sous forme de propriétés. Pour les satisfaire, il faut souvent transformer le graphe initial. On sera alors attentif à fournir des algorithmes qui minimisent son altération lors du processus d'anonymisation. Dans cette thèse nous mesurons l'altération subie par le nombre de transformations élémentaires effectuées sur le graphe. Nous proposons d'étudier la minimisation du nombre de rotations d'arêtes nécessaires sur un graphe afin que celui-ci respecte une propriété d'anonymisation donnée dans différents cas de figure. Des travaux empiriques ont déjà été publiés et nous nous sommes attachés à les poursuivre sous l'angle de la complexité et de l'approximation. L'utilisation de ce nouvel opérateur a permis des résultats théoriques négatifs ainsi que positifs, l'ensemble étant plus encourageant que les résultats sur les précédents opérateurs.
Une communauté est un groupe social partageant un espace, des biens ou des intérêts communs. Cela se reflète sur le graphe par des fortes connexions entre ses individus. Il existe de nombreuses manières de caractériser les communautés et nous nous sommes concentrés sur les définitions basées sur l'optimisation d'une fonction de densité. Plus précisément nous étudions un problème de partition en sous-graphes denses où l'on cherche à maximiser une fonction de densité globale. Nous poursuivons les premiers travaux entrepris sur ce sujet et proposons des résultats plus approfondis.

MOTS CLÉS

Complexité, approximation, graphe, anonymisation, communauté

ABSTRACT

In this thesis we study the algorithmic complexity of optimization problems related to the study of social networks. A social network can be modeled by a graph in which the vertices represent the members and the edges their relations. We are interested in two different problems, graph anonymization and community detection.
Graph anonymization has already been studied and many formalizations of the concept have been proposed in the form of properties. To satisfy them, it is often necessary to transform the initial graph. We will then be careful to provide algorithms that minimize its alteration during the anonymization process. In this thesis we measure the alteration by the number of elementary transformations performed on the graph. We propose to study the minimization of the number of edge rotations required on a graph in order for it to respect a given anonymization property in different cases. Empirical work has already been published and we have been working on it from the point of view of complexity and approximation. The use of this new operator has led to negative as well as positive theoretical results, all of them being more encouraging than the results on the previous operators.
A community is a social group sharing a common space, goods or interests. This is reflected on the graph by strong connections between its individuals. There are many ways to characterize communities and we focus on definitions based on the optimization of a density function. More precisely, we study a problem of partitioning into dense subgraphs where one seeks to maximize a global density function. We continue the first work undertaken on this topic and propose more thorough results.

KEYWORDS

Complexity, approximation, graph, anonymization, community