Subgroup Discovery for Structured Target Concepts

A dissertation submitted towards the degree of
Doctor of Natural Sciences (Dr. rer. nat.)

of the Faculty of
Mathematics and Computer Science

of Saarland University

on Saarbrücken, 2023
by Janis Kalofolias

Tag des Kolloquiums:
8. Dezember 2022

Dekan der Fakultät:
Prof. Dr. Jürgen Steimle

Prüfungsaussschuss:

Vorsitzender der Prüfungsausschusses:
Prof. Raimund Seidel

Berischterstatter:
Prof. Dr. Jilles Vreeken
Prof. Dr. Gerhard Weikum
Prof. Dr. Peter Flach

Wissenschaftlicher Mitarbeiter:
Dr. Philipp Müller

Acknowledgements

As the first, sole informal—and perhaps at times awkward—step, I want to thank everyone who played a crucial role in this thesis: a seemingly Titanean, yet hopefully not titanic, and at times clearly Sisyphean work. Remember, this step is always harder to write than read.

First, I want to thank Mario Boley, who showed me the first steps in academia and instilled me with his ideals on scientific work, research process, and a great amount of inspiration. Our coexistence was short, but I thank you for the impactful (pun intended) care and insight!

I want to extend my grattitude to my advisor, Jilles, who believed in me when it was harder than now! Appart from your scientific contribution, I value our insightful discussions and teaching for the real workings of academia. I appreciate your rare soft skills that I believe should not be taken for granted in academia, and I plan to nurture this voice in the back of my mind, that reminds: see the forest for the trees.

Ευχαριστώ τους γονείς μου, μητέρα και πατέρα, που ήσασταν δίπλα μου όταν έπρεπε, ακόμη κι όταν δεν ήμουν εγώ εκεί—δεδομένης και της απόστασης. Ευχαριστώ τον αδερφό μου! Παραμένεις και θα παραμείνεις σημαντικός και καίριος! Ευχαριστώ τους συνοδοιπόρους μου στον άθλο αυτό, Ε.Ν.Τ. και Π. Μ.: ήσασταν απαραίτητοι.

I thank my friends and coworkers! Our discussions have helped this work, in ways that were not always explicit or clear to any of us. You know who you are! I am thinking of you even without naming you and will keep your identity a cherished mystery.

¡Orya, gracias a ti! Me apoyastes incluso cuando yo no podía, de una manera que no esperaba.

Abstract

The main object of study in this thesis is subgroup discovery, a theoretical framework for finding subgroups in data—i.e., named sub-populations—whose behaviour with respect to a specified target concept is exceptional when compared to the rest of the dataset. This is a powerful tool that conveys crucial information to a human audience, but despite past advances has been limited to simple target concepts. In this work we propose algorithms that bring this framework to novel application domains. We introduce the concept of representative subgroups, which we use not only to ensure the fairness of a sub-population with regard to a sensitive trait, such as race or gender, but also to go beyond known trends in the data. For entities with additional relational information that can be encoded as a graph, we introduce a novel measure of robust connectedness which improves on established alternative measures of density; we then provide a method that uses this measure to discover which named sub-populations are more well-connected. Our contributions within subgroup discovery crescent with the introduction of kernelised subgroup discovery: a novel framework that enables the discovery of subgroups on i.i.d. target concepts with virtually any kind of structure. Importantly, our framework additionally provides a concrete and efficient tool that works out-of-the-box without any modification, apart from specifying the Gramian of a positive definite kernel. To use within kernelised subgroup discovery, but also on any other kind of kernel method, we additionally introduce a novel random walk graph kernel. Our kernel allows the fine tuning of the alignment between the vertices of the two compared graphs, during the count of the random walks, while we also propose meaningful structure-aware vertex labels to utilise this new capability. With these contributions we thoroughly extend the applicability of subgroup discovery and ultimately re-define it as a kernel method.

Zusammenfassung

Der Hauptgegenstand dieser Arbeit ist die Subgruppenentdeckung (Subgroup Discovery), ein theoretischer Rahmen für das Auffinden von Subgruppen in Daten—d. h. benannte Teilpopulationen—deren Verhalten in Bezug auf ein bestimmtes Targetkonzept im Vergleich zum Rest des Datensatzes außergewöhnlich ist. Es handelt sich hierbei um ein leistungsfähiges Instrument, das einem menschlichen Publikum wichtige Informationen vermittelt. Allerdings ist es trotz bisherigen Fortschritte auf einfache Targetkonzepte beschränkt. In dieser Arbeit schlagen wir Algorithmen vor, die diesen Rahmen auf neuartige Anwendungsbereiche übertragen. Wir führen das Konzept der repräsentativen Untergruppen ein, mit dem wir nicht nur die Fairness einer Teilpopulation in Bezug auf ein sensibles Merkmal wie Rasse oder Geschlecht sicherstellen, sondern auch über bekannte Trends in den Daten hinausgehen können. Für Entitäten mit zusätzlicher relationalen Information, die als Graph kodiert werden kann, führen wir ein neuartiges Maß für robuste Verbundenheit ein, das die etablierten alternativen Dichtemaße verbessert; anschließend stellen wir eine Methode bereit, die dieses Maß verwendet, um herauszufinden, welche benannte Teilpopulationen besser verbunden sind. Unsere Beiträge in diesem Rahmen gipfeln in der Einführung der kernelisierten Subgruppenentdeckung: ein neuartiger Rahmen, der die Entdeckung von Subgruppen für u.i.v. Targetkonzepten mit praktisch jeder Art von Struktur ermöglicht. Wichtigerweise, unser Rahmen bereitstellt zusätzlich ein konkretes und effizientes Werkzeug, das ohne jegliche Modifikation funktioniert, abgesehen von der Angabe des Gramian eines positiv definitiven Kernels. Für den Einsatz innerhalb der kernelisierten Subgruppentdeckung, aber auch für jede andere Art von Kernel-Methode, führen wir zusätzlich einen neuartigen Random-Walk-Graph-Kernel ein. Unser Kernel ermöglicht die Feinabstimmung der Ausrichtung zwischen den Eckpunkten der beiden unter-Vergleich-gestelltenen Graphen während der Zählung der Random Walks, während wir auch sinnvolle strukturbewusste Vertex-Labels vorschlagen, um diese neue Fähigkeit zu nutzen. Mit diesen Beiträgen erweitern wir die Anwendbarkeit der Subgruppentdeckung gründlich und definieren wir sie im Endeffekt als Kernel-Methode neu.

Contents

1 Introduction

> Where is the wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?
>
> *(T. S. Elliot, The Rock)*

We live in the age of information, where every aspect of our life leaves a considerable trail of data [RGR17; RGR18], which becomes a valuable resource for the understanding the processes that generate them. The modern techniques of acquisition, however, easily produce an overwhelming amount of information that quickly becomes inhibitive for the unaided human to process. It is therefore necessary for the humans to employ tools that can automatically search for "nuggets" of important information within this sheer amount of data, and then present them in a form that is easily understandable to the human. This comes in contrast with the goals of standard machine learning, where the aim is rather to replace the human in the decision making process altogether. What is more, then the decision taking is instead left upon opaque, black-box models, which cannot be easily inspected by humans to verify their correctness or gain insight on the data. On the contrary, in this dissertation we put forward a paradigm where the machine takes up the role of an intermediary whose goal is to describe important parts of the underlying processes to the humans, thereby helping them take informed decisions themselves.

Consider, for example, that we study the behaviour of a newly identified disease in order to find out which sub-populations of patients are more vulnerable to this disease. Having this information could help us target research efforts on the vulnerable patients, invest funds to create more supportive infrastructure like health centres and hospitals, or simply enable us to educate the affected groups. To identify such sub-populations, we could first create and study a sample of the numerous patients that were tested positive for the pathogen. For each patient we would also record important demographic information (e.g., sex, age, race, etc), medical history (e.g., years smoking, possible comorbidities, etc) and other relevant information (e.g., date of admission); we can also measure the vulnerability of a patient by

tracking either the number of days until their full recovery or—wherever more appropriate—whether they survived an infection. This kind of collected data contains information not only about the usual behaviour of the pathogen, but also allows to pinpoint sub-populations which stand out with respect to this usual behaviour. Therefore, this dataset could allow an automated intermediary to test a collection of several hypotheses, and then present to the human the one that answers best the original question: What profile describes the patients which are unusually more vulnerable to the pathogen?

An exemplary answer to this question was given while studying data gathered from British hospitals early on during the SARS-CoV-2 pandemic. Platt and Warwick [PW20] report: "*Patients of the Bangladeshi minority that live in London exhibit* 2 *times higher fatality rate than normal*". Although there is certainly the potential to use this description as an interpretable classifier to predict higher fatality rates, the true value of this result lies in its inherent intelligibility. That is, this description clearly outlines a vulnerable sub-population of patients, while it additionally interprets the effect of belonging to this sub-population. Such a description can therefore be readily conveyed to a domain expert—that need not be a mathematical specialist—and is thereby actionable: The policy maker among them can increase facility availability to this minority, the medical personnel can treat those patients more urgently, while even the general public can be more cautious when interacting with their more vulnerable fellow citizens. At the same time, such a result can be a valuable tool in finding potential directions of future research.

Indeed, a later study [ZP20] identified a small region in the Homo Sapiens genome that poses a major genetic risk factor for this disease. This haplotype was highly prevalent in Asians, among which with the highest occurrence (63% single and 13% homozygous) in people of Bangladeshi genetic origin, thus explaining the previously observed anomaly in the fatality data. This case study vividly demonstrates a desirable trait of this intermediary algorithm: that its output must be an "intelligible" description. Such is a description whose meaning can be easily understood by a human and it can be easily conveyed from one person to another, and thus action can be taken based on it. This notion is different than what both terms of 'interpretable' and 'explainable' have come to define [RCC+22] in the context of modern machine learning. An **interpretable method** refers to one which makes it possible to discern the mechanism that led from the input to the result, that is, it is transparent as to *how* a decision or prediction was made; the opposite term is un-interpretable, which is synonymous to a '**black-box**' method (or model). On the other hand, the goal of **explainable methods** is to provide

post-hoc an intelligible result as an alternative to an otherwise black-box model. From this perspective, an intelligible description, such as the above, can also function as an interpretable classification model: the decision of whether a patient is high-risk is based on whether he is Bangladeshi and Londonese. We hence require that an intermediary algorithm must exhibit this useful trait.

Requirement R1. *Find intelligible descriptions of important parts of the data.*

While seemingly easy to fulfil, this requirement has the profound implication that in our derivation of an intermediary algorithm we have to let go of one of the most common mechanisms that power the majority of standard machine learning methods: that of continuous optimisation. In contrast to using a set of continuous parameters to combine traits, satisfying Requirement R1 requires to specify a subset of the data as a sub-population, which is a typical combinatorial problem. What complicates matters even more—and in fact substantially—is that it is necessary to search exclusively among those subsets that correspond to describable sub-populations, since it is usually impossible to find an exact description from an arbitrary subset of patients. Thus, our task is placed within the family of constrained combinatorial problems, which comprises a notoriously hard family of optimisation problems.

We henceforth refer to any description that corresponds to an interesting sub-population as a **subgroup**; this leaves as a natural next step to specify what renders one subgroup more interesting than another. As previously motivated, there is valuable information in those particular sub-populations whose behaviour is deviating from that of the whole population. Such a behaviour was in our previous example the exceptionally high vulnerability of the described patients to the pathogen. This establishes that the desired output of the algorithm we are looking for is in the form of a particular subgroup, and more specifically one that refers to a sub-population with a locally deviating behaviour. From this perspective, our task can be described as named anomalous subset selection, since we require of our intermediary algorithm to find an outstanding sub-population with a description: an outstanding subgroup.

Requirement R2. *Find a subgroup that exhibits exceptional behaviour.*

In fact, we might even advance this requirement one step further. To this end, note that Requirement R2 implies the reasonable assumption that there exists a well-defined measure of how exceptional a subgroup is, or, in general,

an appropriate quality measure for a subgroup. For instance, in our example we used as such a measure the ratio of fatality rate in the subgroup compared to the fatality rate over all patients. Once such a concrete measure is specified, we can further define an **optimal subgroup** with respect to this measure: the one that maximises the specified measure. Beside specifying the optimal subgroup, having such a measure further allows our algorithm to provide quality guarantees also for the sub-optimal subgroups. In other words, we can further distinguish different candidates for our desired intermediary algorithm based not only on the quality of their results, but also on the certainty on how good these are. Namely, these criteria are i) whether they provide an optimal subgroup, ii) how efficiently they compute it, or iii) whether they can provide a bound of how close the quality of their subgroup is to the optimal.

Each of these quality criteria for our automated intermediary algorithm has its own merit. Looking past the obvious superiority of the optimal subgroup, it is also desirable to provide optimality bounds for sub-optimal results. Consider, for instance, that finding the optimal subgroup requires prohibitive resources, in which case we would be content with a sub-optimal subgroup—that would be cheaper to discover—as long as it comes with a certificate of how close its quality is to the optimal one. In fact, we can easily construe an example where it might be well preferable to employ an algorithm that would yield on average lower quality subgroups, but equip them with such a certificate of quality, than to use some heuristic algorithm that may on average provide higher quality subgroups albeit never with quality guarantees. The output of the latter, non-exact algorithm, would be akin to an educated guess and would therefore convey little to no information on how much higher the quality of the optimal subgroup is.

This lack of guarantees of the search algorithm can be a substantial disadvantage. As we recall, the optimal subgroup in our example is the Bangladeshi minority in London, which is the most vulnerable among all others; this knowledge justifies the investment of valuable resources to research this peculiarity or to invest available funds in infrastructure that would support this sub-population. On the other hand, if we used a non-optimal algorithm, it would be most likely that it would inform us of another sub-population that may be more vulnerable than usual, however to a possibly much lesser degree. This can deteriorate the objectivity, actionability, trustworthiness and general usefulness of the result in several ways. First and foremost, it leads to the potential of wasting valuable resources in sub-optimal causes. Additionally, it does not inform of the potential sub-optimality of the discovered result, which, in turn, hurts the trustworthiness of the this

result. Indeed, if a policy maker were to act on this information, it would be difficult to argue that this favouring of the discovered sub-population follows exclusively as the result of a sub-optimal algorithm, instead of an arbitrarily partial choice on behalf of the policy maker. In fact, one could mistrustfully—or even legitimately—claim that the policy maker had run this sub-optimal algorithm several times—or alternatively had tried many sub-optimal algorithms—to then pick out of all results the one most appealing to this actor.

From a different perspective, an algorithm that provides quality guarantees can also be used to rule out the existence of a deviating sub-population. Although not frequent, it is realistically possible that, within the available sample of the population, no evidence can be found that a (significantly) exceptional sub-population exists—at least among the sub-populations that we can described using the available data. In fact, whenever an exact algorithm terminates without finding any proof of exceptionality, it ensures exactly of this situation. Such a result is actionable in itself: in our example of patients, it would then render acceptable to relocate available funds from supporting vulnerable populations to other tasks, or to redirect research time on more promising directions. In contrast, when a heuristic algorithm fails to find any acceptable subgroup, we cannot distinguish between i) the non-existence of a sub-population of interest, ii) the case where a sub-population does exist but requires running the algorithm for a longer time, or even worse, iii) that a sub-population exists but lies outside the reach of the algorithm we employed. In fact, we will use one of our exact algorithms in a gently more elaborate setting applied on an extensive dataset of committed crimes (see Section 3.3.1) to provide evidence that "*no race is more violent when it comes to murdering young women*".

This desire for optimality guarantees calls for the use of an exact algorithm, i.e., one that can either provide the optimal result, or alternatively provide bounds on how lower the quality of the provided subgroup compares to the optimal one.

Requirement R3. *The result must either be optimal or come with quality bounds.*

Of course, exactness comes at a cost. More specifically, this requires the use of carefully developed bounds that allow for the efficient optimisation of difficult combinatorial problems, often proven to be NP-optimisation ones. Nevertheless, exactly due to the benefits of exact solvers, in this work we focus on the optimal solution of this task.

1.1 Standard Machine Learning Fails our Requirements

In view of these requirements—and in particular with respect to Requirements R1 and R2—standard machine learning (ML) tools[1] fail to provide a solution to our needs. First and foremost, their goal is not to aid humans in taking decisions, but instead to supplant them in the decision making process altogether. At first sight, when vast amounts of data are provided to standard ML tools, they do yield models able to make accurate predictions that in specific tasks can match—or even exceed—the performance of humans [Ste17a; BGB+19]. Nevertheless, we should not always take the convenience of delegating decisions to the machine lightly, as their abilities often fall short of that of a human, which can therefore not be replaced. This is the case particularly in tasks that are not well-defined, involve subtleties that are hard-to-describe, might be susceptible to biases, or can have morally grave repercussions [Bod21].

A demonstrative case study of such shortcomings is the Google Photos platform: a service that allows users to upload pictures and offers a feature to tag them based on their content. This feature became the cause of turmoil when it wrongly tagged African-American users as gorillas [Bar15]—a pejorative remark that a human would readily recognise as a common racial slur. This was due to a mis-classification error that was unable to be fixed even after years of advances in models and further collection of data, and leaving as only viable workaround the complete removal of this and several other related offending tags [Sim18]. Even more recently, several issues of this kind appeared once again in a similar platform [Mac21; Moo21]. Likewise, broad application domains can also be affected, with repercussions that can go well beyond amusing or embarrassing. Mis-classifications in tasks like facial recognition can hurt the self-identification and dignity of individuals and even threaten their freedom and livelihood [Wae22]; such cases have been repeatedly reported when innocent individuals were falsely identified as suspects and wrongly incriminated [And20; Fis20]. What is additionally alarming is that such tools often disproportionately mis-classify non-white faces, for instance due to systematic differences in skin tone or due to lighting conditions optimised only for Caucasians [Naj20].

Perhaps even more alarmingly, this disparity in mis-classification of sus-

[1] As such we consider the popular tools that solve the standard machine learning tasks of supervised and unsupervised learning. Notable examples are linear and logistic regression, classification and regression trees, support vector machines, or, for the unsupervised case, expectation maximisation and k-means. In the broader sense, this category also contains the general probabilistic models, neural networks, and a plethora of other similarly powerful tools [HTF09].

ceptible minorities is an increasingly occurring phenomenon that has been attributed to unavoidable, inherent biases in the originating data. This brings attention to yet another source of hard-to-identify systemic errors [ALM+16] that the machines are susceptible to. As an example, we may consider a popular automated system used in administering justice for the prediction of recidivism[2] that has repeatedly come under scrutiny for its reportedly racially-preferential[3] decisions [Mes21]. All those issues of standard ML tools expose as a misconception the popular belief that the solution is simply a matter of more data, more powerful models, or longer training, and instead raise the point that "artificial intelligence is neither artificial nor intelligent" [Cor21].

Of course, this is not to say that humans are insusceptible to mistakes, and neither to compare the average accuracy between human and machine. Instead, it demonstrates that automated tools lack skills that are necessary for the decision making in certain sensitive tasks, even after we are willing to assume that only few prediction errors will be made. Therefore, in these tasks the machines should not be entrusted with replacing the human in the decision making process. In fact, systems used for hiring employees for open positions are subject to state regulation [Lai21], while the ethics of artificial intelligence in deciding the course of patients has come into question, especially in conditions of humanely allocating limited resources [Wet21]. Importantly, this reluctance to trust such models is further exacerbated by their high complexity, which renders them akin to opaque black boxes. That is, regardless of how accurate the predictions of these models can be on average, a human audience cannot understand why each particular prediction was made individually, and much less to understand which sub-population of individuals is favoured or disfavoured by the model.

Despite undesired, this obscurity of black-box models is neither beneficial not unavoidable. Contrary to this, the promoters of opaque predictive tools often defend the obscurity of their models[4] by attributing it to the purported complexity of the task [RWC20]; this task is then falsely claimed to not admit a highly performing yet simple enough model: one that can be scrutinised by

[2] *recidivist* (noun): a criminal who continues to commit crimes even after they have been punished. Cambridge English Dictionary, 2022.

[3] The COMPASS tool used in 46 of the United States reportedly "incorrectly labelled Black defendants as "high-risk" at twice the rate as white defendants" [ALM+16]. Interestingly, this is not an artefact of model mis-classification, but instead stems from the inherent bias in the collected training data.

[4] It is worth noting that, the creators of such models are incentivised to treat them as proprietary technology and have great interest in keeping them opaque as a trade secret. On the other hand, judges and prosecutors prefer using transparent and comprehensible models whenever possible [Rud15].

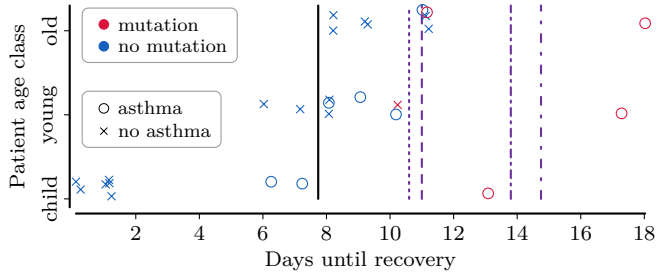
humans, provide intelligible predictions, and ultimately be trusted [RR19]. In fact, recent evidence disproves this rationale [Rud19]; as a demonstrative counter-example, in the task of criminal risk prediction, a simple description of the high-risk population has been demonstrated that can be used to classify the potential offender as a recidivist or not, and with an accuracy that is on par with other state-of-the-art methods [ALA+18]. Using a method outside the standard ML tool-set, the authors show that offenders belonging to a given sub-population are likely to be rearrested within 2 years of their evaluation. This populations accepts the following intelligible description:

i) "persons with > 3 prior crimes",
ii) "males between 18–20 years old", or
iii) "persons with 2 or 3 prior crimes and between 21–23 years old".

This description makes it easy for a human not only to apply the implied model, but also to verify it is not riddled with systematic biases that could lead to gross mistakes and potentially affect innocent individuals profoundly. This not only questions the superiority in terms of accuracy of standard ML models, but further showcases the importance of developing models that provide an intelligible description to the human audience.

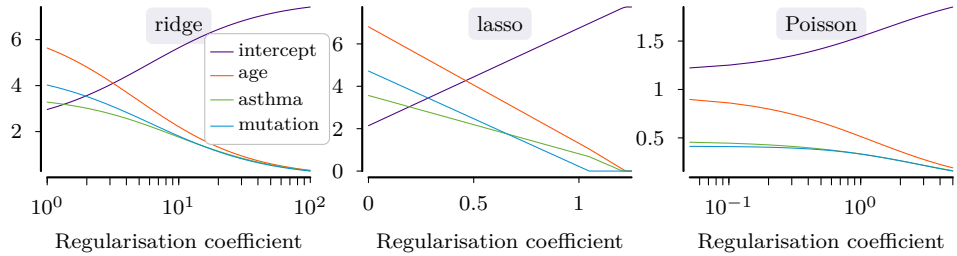
From this perspective, standard ML tools fail to provide a satisfactory insight on these sub-populations of interest, often without offering any benefit over methods which provide intelligible results. Specifically for the simpler among the standard ML tools, a frequently used yet often failing workaround [Lip18] toward getting insight on the corresponding models is to try interpreting the model parameters. To demonstrate this process we may study the toy dataset depicted in Fig. 1.1a that is largely inspired by our running example of patients. Here, we measure the patient susceptibility in number of days until recovery, which in turn depends on three traits of the patient: i) their `age` class, ii) whether they have `asthma`, or iii) carry a `mutation` that introduces susceptibility to the pathogen. Using just a short glance in the visualised data, the human eye can quickly discern sub-populations of patients with high susceptibility: in contrast to the general average of 7.7 days, patients with `mutation` (red marks) need 13.8 days on average, those with `asthma` (circular marks) 10.6, while those with both traits need 14.7 days (!).

On the other hand, observing the coefficients of simple standard ML models demonstrates that this approach provides neither an intelligible description, nor the necessary insight. To show this, we depict in Fig. 1.1b the continuous coefficients of three representative ML regression models that we trained on the above data. These models are arguably among the easiest to interpret, while retaining substantial flexibility: that of ridge regression,

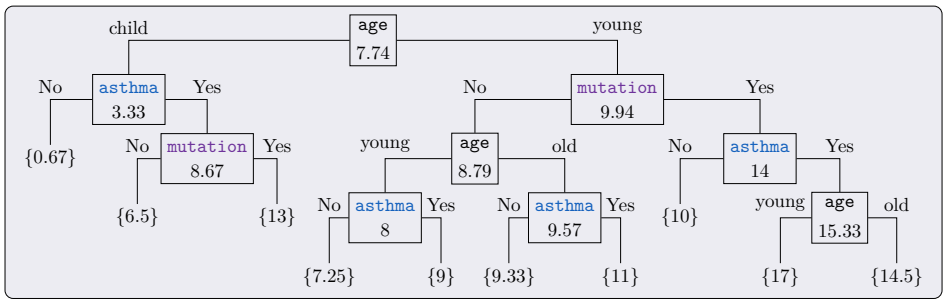


Sub-Population	Days (avg)
all patients	7.7
old patients	11.0
asthma	10.6
mutation	13.8
asthma & mutation	14.7

(a) Patient key traits and their susceptibility, alongside subgroups of interest.



(b) Model coefficients of standard ML tools fitted on the above data for a varying degree of regularisation. Although these models strive to remain simple enough for the sake of interpretability, their coefficients provide neither an intelligible description nor a good indication of important subgroups. Indeed, even though in these models the coefficients of both `asthma` and `mutation` are systematically overshadowed by the less important `age`, the first two features both form important subgroups, and even more so their combination.



(c) The complete regression tree fitted on the data, listing in each node the average number of days. The local effect of both features `asthma` and `mutation` is overtaken by the larger, global effect of `age`; this delays the selection of these former, important subgroup features and misleadingly distributes their effect under multiple branches of the latter, less relevant, and preventing the discovery of the desired subgroups.

Figure 1.1 [Shortcomings of Standard ML]: Toy dataset demonstrating common mechanisms that cause standard ML tools to fail our requirements. A human (and a subgroup discovery algorithm) can correctly identify important subgroups, such as "patients with mutation" or "patients with asthma and mutation". Standard ML tools, however, can not only miss the important, local subgroups in their attempt to model the entire dataset, but also yield much less intelligible results—if at all.

lasso, and also the generalised Poisson linear model that is the model-of-choice for survival data [CRS+12], and therefore closely matches the data at hand. However, after inspecting these models, the closer to a description that we can get—say, from ridge regression ($\alpha = 1$)—is a result of the form: $2.59 + 6.17x_{\texttt{age}} + 3.42x_{\texttt{asthma}} + 4.34x_{\texttt{mutation}}$, which is arguably far less intelligible than the desired subgroup: "*patients with* ***asthma*** *and* ***mutation***". Even worse, in realistically sized data, the number of non-zero coefficients can be prohibitively high for a human to process, or can be misleading due to unaccounted-for dependencies between the covariates, or have negative signs, all of which prevent the derivation of a useful description. In realistic applications, similar complexity is also encountered by even simpler methods that discount their flexibility in favour of their interpretability, such as regression trees (see Fig. 1.1c). Of course, going in the opposite direction of more powerful models—as computed by kernel methods, support vector machines, or neural networks—only exacerbates the difficulty of interpreting the model coefficients. Then one needs to resort to further tricks, as is for instance the retro-fitting of one of the above simpler models locally around a particular prediction [RSG16]. Unsurprisingly, in addition to only being accurate around just a single prediction, these methods still yield standard ML models, with all their disadvantages.

Even past the difficulty of describing the results of standard ML tools, there is yet another crucial front where these tools fail to deliver: the standard ML models have a global character versus the local behaviour that we desire. More precisely, these ML methods focus on modelling the entire dataset, which comes in stark contrast with a key goal of our task: that of finding an exceptional subgroup. By definition, such a subgroup describes the behaviour of a sub-population that is essentially a small, local part of the data, the behaviour of which must stand out from that of the entire population. That means that, if we were to fit some global model on the population, it would—by definition—not be a good fit for the sub-population corresponding to any high quality subgroup.

As a result, these global models often fail to reflect the effect of a small yet important sub-population in the data; indeed, this relevant local anomaly might only have a small effect when modelling the entire dataset, or its effect might be overshadowed by other features which impact a wider part of the data. This artefact is already evident in the coefficients of all of the standard ML models that we encountered in Fig. 1.1b: Here, the `age` feature describes a larger part of the data when compared to that of both `mutation` and `asthma`, which causes the coefficient of the former feature to misleadingly overshadow the coefficients of the other two. This goes against both our

intuition and desire, which dictate that `mutation` and `asthma` describe both a small sub-population which exhibits an important deviation from the usual duration of sickness, and therefore constitute high quality subgroups. Importantly, the same effect also appears in the case of other more easily interpretable models; among them, arguably the most characteristic are regression trees, which we train on our data and demonstrate in Fig. 1.1c. Here, the trait `age` similarly overshadows the other two, which prevents the discovery of the desired subgroups. As a result, we only encounter these traits combined with that of `age`, and still with a significantly lower importance; for instance, on the right of the tree we encounter the sub-population described as "*not* ***young*** *with* ***mutation***", which in contrast to the sub-population of our desired subgroup "*patients with* ***mutation***" exhibits a mean of 10 days instead of 13.8.

All in all, these arguments do not call for a complete banning of the use of standard machine learning methods for certain sensitive tasks; instead, they call for new automated tools that can extend the human intuition on real-world data with complexities that readily overwhelm the human eye. In other words, they make a case for a new paradigm of "augmented intelligence" [CK21][5], where automated tools work to empower humans to make decisions. In fact, the main study of this thesis is the very premise of automatic tools that act as intermediary between data and the human audience, by describing important sub-populations in the data.

1.2 Subgroup Discovery to the Rescue

Fortunately, there is a powerful theoretical framework whose goal is exactly to find intelligible descriptions of exceptional sub-populations. This framework is aptly called subgroup discovery (SD) [Klö96; Wro97; Web01; LKF+04], and has proven its versatility through its very broad applications [Hel16]. Important applications have been demonstrated in material science research [GBV+17; SBG+20], in political science [BCL+20] and election results [GBK10] in particular, as well as medicine and bioinformatics [MRS+09; HCG+11; AHM+21], just to name a few.

Our goal within this framework is akin to the supervised local discovery of subgroups that are deemed interesting according to certain criteria, as

[5] The second author, Garry Kasparov, secured the chess grand master title in November 1985, but became widely known to the general public after losing a tournament to the IBM supercomputer *Deep Blue*, in 1998. The authors note that in a recent worldwide competition, where apart from human players it was also supercomputers and mixed teams competing, the winner was a couple of amateur human players aided by 3 averagely-powered computers.

captured by a well-defined objective function. More specifically, subgroup discovery can be formally seen as the problem of optimising the chosen objective function over a set of allowed subgroups, which is the constraint enforced to fulfil Requirement R1. More specifically, the allowed subgroups are those subsets of the dataset with descriptions that typically consist of simple conditions that must be satisfied together, thus forming increasingly more specific expressions [Klö02]. In our example, such conditions were: "*citizen of London*" and "*of Bangladeshi origin*", which combined described the subpopulation of interest. This choice at once produces intelligible descriptions and allows for a convenient way to efficiently optimise over the set of all subgroups. Requirement R2 is reflected in the choice of the objective function, which quickly makes the full versatility of this framework apparent. Indeed, during its extensive study in the literature for about 3 decades many objective functions have been proposed, and from very different perspectives throughout the years.

As a first categorisation, the existing subgroup discovery methods can be broadly distinguished based on the objective function [Atz15] that they propose into *user-centric* and *data-centric*[6]. The user-centric methods aim at choosing subgroups that are interesting according to some model of user preferences. Notable examples of these methods use concepts from information theory, for instance to look for subgroups that are surprising according to a model of prior user beliefs. Such models have been derived according to the maximum entropy framework [LKD+18], which tries to make no further assumptions beyond the ones encoded in the adopted belief system [Jay82]. Other methods use the Minimum Description Length principle [vLK12] that corresponds to a particular subjective prior on how the data have been generated, in a way that favours traits widely assumed to be appropriate for humans, namely the simplicity of its representation [Ris78; VLS11].

In contrast, data-centric methods use measures that assess the fitness of each subgroup based solely on the **entities** of the given dataset, as we call the objects of study within the dataset (e.g., patients). Typically, these methods use principles from statistical theory to derive objective functions that are based on appropriate test statistics or distribution distances. In this setting, the aim is to find a describable sub-population within which a specified random **target variable** exhibits the most exceptional behaviour when compared to an appropriate null model for this variable; this Null

[6] These terms are synonymous to the originally used "subjective" and "objective" quality measures [Atz15], respectively; we here use their synonyms to avoid confusion with our adoption of the term "objective function" for the quality measure of each subset.

model is derived from an appropriate sample of the true population, which is usually the entire dataset. Hybrid methods have also been proposed, for instance using the additional pervasive assumption that users prefer more succinct subgroup descriptions [vLee10], which, to an extent, we also take into consideration in our own work. Importantly, data-centric methods can also serve as a key first step to select a set of subgroups, which are subsequently filtered according to user-centric criteria, for instance to yield diverse results [vLK12]. Hence, in this work we focus only on the data-centric methods.

Another important distinction of both classes of methods[7] is the nature of the target variable. Most methods assume this to be a scalar value that is directly provided within the dataset as an attribute of the studied entities [HCG+11]. These can take on Boolean [KLG+09], categorical [SKF+16] or numerical values [GR09; BGG+17; LKD+18; PGB+21], as were in our example the Boolean tracking whether the patient died, or the numerical time until their complete recovery. Each of these cases can define a rather broad family of works; for instance, *contrastive set mining* has been shown to be the special case of subgroup discovery with Boolean targets [KLG+09], while categorical targets are a natural multi-class extension of Boolean ones [SKF+16].

From this setting onward the target variables give way to more general **target concepts**, which are targets implicitly derived from the properties of each entity of the entire subset. We can decry two very similar directions with blurred boundaries to derive target concepts.

The first direction is the broad family of exceptional model mining (EMM) [LFK08; DFK16], which was primarily motivated by the case of multivariate target variables. Here, a particular model of data is chosen, after which the objective function measures the difference of the model parameters between i) those fitted on the entire dataset and ii) those fitted on the candidate subgroup . This model may take many forms, including a linear classifier, a simple decision tree [LFK08], or even a probabilistic Bayesian model [DKF+10; SFK15; SBD+22]; for each model different statistics have been proposed, such us a z-scores [LFK08] or a likelihood ratio [SFK15]. The EMM paradigm can also be seen as using a target concept that constitutes a random variable derived from a set of entities, for instance that of the correlation coefficient between two scalar attributes, and then using an appropriate statistic for this derived random variable. This introduces

[7] Although the distinction based on the target variable applies to both user- and data-centric methods, when we refer to the target variable as a random quantity the focus is on the latter category.

the second direction, in which the target concept is not an explicitly listed attribute of the dataset, but is instead derived based on properties of the entities. Perhaps the most characteristic examples are given by the case of named community detection [ADM16; LKD+18], where the target concept is the community-like behaviour of the sub-graph. We will encounter another example from the task of named dense subgraph mining as part of the contributions of this work, later on in Chapter 4.

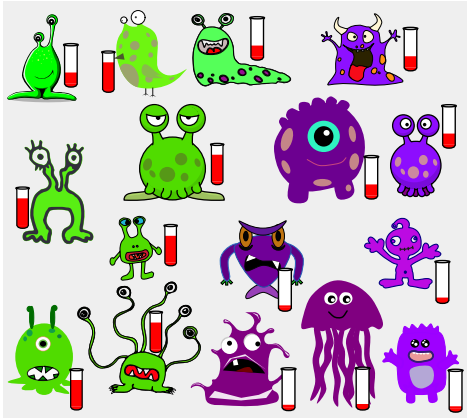
1.3 Subgroup Discovery for Structured Target Concepts

Despite the strong promises of this framework, the state-of-the-art in subgroup discovery still leaves a lot to desire, as its applicability is limited on target concepts with trivial or no structure.

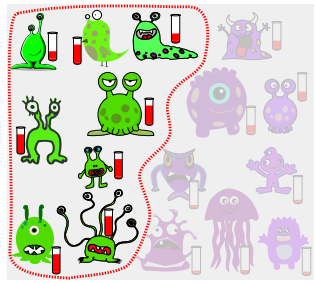
Existing methods are oblivious, for example, to previously discovered sub-populations, making subsequent discoveries highly redundant. At the same time, existing methods do not respect sensitive variables, such as gender or race, that we already witnessed to be often subject to inherent biases that seep in the data. These methods can therefore yield subgroups which misrepresent the sensitive population, a situation that can easily lead to unfair treatment. In this work we will study ways to ensure that a subgroup remains representative, so that it represents novel discoveries, that extend prior knowledge or gross trends in the data. We will also show how our approach can be further used to fairly describe sub-populations with sensitive traits, ensuring no minority is misrepresented.

Another great challenge arises when the population has additional structure in the form of relations between the studied entities. A plethora of works treat this problem as the related community detection one, however they either provide no descriptions, are non-exact, or use degree-based measures that fall short of capturing robust connectedness. In Chapter 4 we first define a measure of robust connectedness that takes into consideration the average number of relations that must be severed for the entities in the subgroup for it to become disconnected and show how to optimise it within the SD framework.

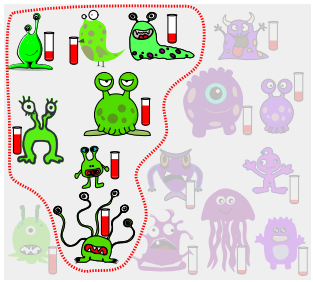
When it comes to exceptional model mining (EMM) methods, despite their flexibility they resort to heuristics to optimise the proposed objective functions, and thus offer neither an optimal answer and neither any quality guarantees. In addition, these methods suggest models that have to be simple enough so that the derived statistics retain their statistical properties, therefore leaving out the powerful non-linearity and expressibility of a positive



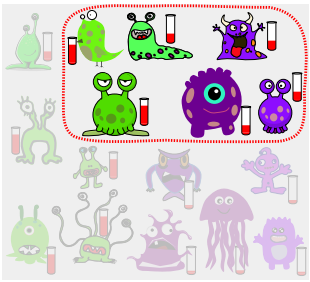
(a) Bacterial species and their dye affinity. The colour indicates the Gram stain result: gram+ species are indicated with green and the gram- with purple, while the affinity of the dye is represented as the level on each test tube.
In this toy example we model a high correlation between gram+ categorisation and dye affinity, as would be the case if we were to study the fluorescent wheat germ agglutinin dye [SCK90]. Here, our goal is to find subgroups that represent both gram+ and gram- species equally.



(b) Naïve description: appear only gram+.



(c) Gram attr. removed: appear almost the same.



(d) Representative query: both types appear equally.

Figure 1.2: Bacterial species (personification) used to study the affinity of a fluorescent dye, beyond its known sensitivity to gram+ species.

definite kernel. As a further downside, the EMM methods often require expensive computation procedures to fit the used models for each candidate subgroup, which can in general quickly become inefficient, despite proposed sampling schemes to amend the issue for certain models [MB14].

In this work we address each of these issues, which we briefly motivate below.

1.3.1 Representative Subgroup Discovery

Precisely due to its remarkable ability to find the most exceptional subpopulation in the data, standard subgroup discovery can be very susceptible to trends or biases within the available sample. Let us consider once more our example of patients, where discovering the Bangladeshi minority to be highly susceptible is normally a very desired response on behalf of the algorithm. That is, the very same trend for Bangladeshi patients to exhibit higher risk to

the pathogen, knowing of which is normally a desired result, can immediately turn into a problem once we already know of its existence. Then, to be able to perform further discoveries we must now be able to steer the algorithm away from this known-to-be-susceptible sub-population and, instead, toward a novel, undiscovered one.

To shed more light in this mechanism that leads the standard subgroup discovery to failure, we remain in the field of medicine and biology, where it is useful to study the response of large samples of bacterial species on novel substances. Let us imagine, for instance, that we study a novel fluorescent dye in order to understand its affinity on numerous samples, each of a different bacterial species. Apart from the affinity of the dye on each sample, we also have available known structural properties for the species, e.g., number of membranes, shape, mechanism of movement, etc, among other characteristics of interest, such as contained organelles, antibacterial resistance, etc. Beside these traits, the dataset also lists the Gram stain [Coi06] categorisation of each species into `gram+` and `gram-`, which is a major first step in bacterial classification. It thus becomes quite plausible that one of the two classes of bacterial species have significantly higher affinity to the studied dye. In fact, this would exactly be the case, if we were using a fluorescent variant of the wheat germ agglutinin dye, to which it is known that the `gram+` bacteria have a substantially higher affinity [SCK90].

As posited in our example, in several cases the researcher already knows of a major trend in the data. Let us introduce as a visualisation aid a toy collection of bacterial species in Fig. 1.2a, gently personified for easier depiction. If typical subgroup discovery were to be performed on this data it would find that the bacteria to which the dye has greater affinity are "*gram+ bacteria*", which is something that we already knew, and is thus uninformative, and therefore undesired. The problem is also not solved by simply removing the trait of `gram+`, shown in Fig. 1.2b. Indeed, invoking the same algorithm on data, from which we remove the gram classification that we want to avoid, we would get for this example: "*bacteria with external eyes and no tentacles*". That is, yet again, we discover almost the same, known fact, just with a different name, as shown in Fig. 1.2c.

Query Q1.a. *What affects bacterial affinity, beyond Gram classification?*

To go beyond what we already know, instead of ignoring the prior information, we instead have to incorporate it in the algorithm, leaning toward a sub-population that equally corresponds to the `gram+` and the `gram-` bacteria. If we were to do so, we would indeed discover the sub-population "*spotted bacteria*", which surpasses—and thereby adds to—our current understanding of the problem.

Interestingly, the same toy example demonstrates how biases in the data can lead to unfair results that can lead to disparate treatment. To see this, let us imagine that each of the entities depicted in Fig. 1.2a is a candidate student, whose previous educational institution lies either in a wealthy country (green) or one not so much (purple), and assume the indicated value to be the amount of extra curricular activities of each student. In granting a scholarship for these students, we would like to represent those with higher engagement in extra-curricular activities, while avoiding partiality toward wealthier countries, not only due to a desire for fairness, but also due to the amount of extra curricular activities being highly affected by the wealth of the originating country. If we simply asked a subgroup discovery algorithm unbridled, it would indicate that the wealthy students were to be preferred, either directly—by explicitly mentioning their originating country—or indirectly—by re-describing the same group of students through other means.

Query Q1.b. *What student profile has high extra-curricular engagement, while equally representing both origins?*

One way to address both Queries Q1.a and Q1.b lies in shifting our answer toward a pre-determined distribution of an additional control property of interest: here, the gram categorisation of the bacteria or the student origins. In Chapter 3 we introduce the notion of representative subgroups and will use it to develop a method that can answer both these questions, and thereby obtain the first method for discovering optimal subgroups that globally representative of a sensitive trait, while remaining locally exceptional. In other words, both above queries can be answered by a method that can solve the following research goal.

Goal G1. *Find the subgroup with the most exceptional target variable whose control property is representative of the desired distribution.*

1.3.2 Entities with Structural Relations

Some information between entities, however, cannot be adequately captured by treating them individually, as we did so far. Indeed, in our study of SARS-CoV-2, we know that the transmission of the virus succeeds after close contact. We can therefore imagine a dataset in which consenting citizens provide their demographic and medical information to an authority, but also allow the use of electronic means to track their geo-location, for instance through telecommunications records, their travel forms during flights, or using a mobile application to track the social events they participated. We

can thus use all this information to infer whether there has been a possible exposure between any two individuals.

From a broader viewpoint, so far we treated the entities involved as independently and identically distributed samples of a statistical population. This assumption however quickly break down when the entities of the population were not independently generated, as is commonly assumed in machine learning, but were rather somehow structurally dependent on each other. This is the case in our example of SARS-CoV-2 patients when we additionally consider their exposure information. Similar kinds of information arise naturally in many aspects of the social human interactions, and being able to discover which clusters of them form well connected sub-populations offers valuable insight in the underlying processes.

Returning to our example, and after having encoded this information as a graph, we can now query the following.

Query Q2. *Which subgroup describes exposed super-spreaders, each of which exposed several other super-spreaders within the same subgroup?*

In other words, we are not simply content with finding a sub-population of individuals which have spread the pathogen to several others, that may or may not belong to the same subgroup. Neither do we look for a sub-population of individuals that simply have a high average of exposures; this could simply be the case if only a few super-spreaders infected a large group of individuals, Instead, we additionally require that each person within the sub-population has spread the pathogen to many others within the same sub-population: each member of the subgroup must be a super-spreader with regard to the same subgroup.

An important case study that is demonstrative for the importance of this additional structural requirement, is the unexpectedly high rate of SARS-CoV-2 contagion in the Gütersloh area of Germany, where more than 2000 citizens were tested positive. Having this information we can imagine a dataset of all the positively tested residents of this area that forms a graph, in which every node is an individual with census attributes such as their occupation, and each edge encodes corresponds to the existence of a possible route of exposure between the connected individuals.

Using such a dataset, an algorithm tasked with Query Q2 would be able to discover the exemplary answer: "*The workers in the refrigerated meat preparation department of the Tönnies meat processing factory in the Gütersloh area were each very highly exposed to each other*". We base this answer on the reported [Nor20] breakout in the aforementioned meat processing factory, whose refrigerated department was operated without

necessary oversight or precaution; namely, the cold air in the refrigerated area was circulated without appropriate filtration, and there were insufficient isolation measures for the overpopulated canteen, where employees were in close contact and without personal protective equipment. This discovery led to a quick first response of measures from the authorities [Jan20] as well as further investigations and subsequent revelations of common malpractices in this industry [Lee20].

Queries like Q2 are highly related to the task of community detection, which is an established tool in network understanding, and often the first step in social network analysis. In this work, however, we not only find a description of the community-like cluster of interest, but we also revisit the conditions that form an interesting cluster of connected individuals. In our example, this would be the requirement that each worker had to be exposed to a minimum number of others in the same subgroup, on average.

We call this measure **robust connectedness**, since it builds on the concept of a minimum amount of connectivity, in that it guarantees the network to remain connected even when an adversary was allowed to remove a given number of arbitrary edges from the network. Thus, the robust connectedness we propose induces a stricter condition than simply measuring the average edge-to-node ratio within the community, or, equivalently, the average node degree of the community, which form the basis of typical measures of connectivity. Indeed, a network consisting of a few super-spreaders and otherwise several other vertices that only connect to the former ones, could easily fall apart and become completely disconnected, once we remove these few super-spreaders, despite having an astonishingly high edge-to-node ratio. The very same network, however, would score significantly lower in terms of robust connectedness than a more "well"-connected network with the same, or even much lower edge-to-node ratio.

Thus, our measure is a more structurally-aware measure of connectedness than typical metrics used in dense subgraph detection, as well as its extension, the community detection. We hence show how to find named sub-populations whose entities are well-connected, which thus reveals a novel type of insight that goes beyond that offered by the simpler requirement of densely connected entities. In essence, answering queries like Q2 can be fulfilled by solving the equivalent research goal.

Goal G2. *Find the subgroup whose entities induce a robustly connected graph.*

1.3.3 A Generalised Subgroup Discovery Framework

Rethinking of the standard subgroup discovery once more, we may already discern two key limitations.

First, in the typical setting, all methods are focused on scalar-valued targets. While there certainly does exist an abundance of applications where the target concept is a scalar variable, this precludes many applications that are as much interesting as important, where we want to study entities whose behaviour cannot be simply summarised by a single scalar. Such applications could be, for example, describing a group of exceptional molecules, stocks, or images.

To provide a more detailed example, let us imagine a collected sample of one or more neuroimaging modalities [WZD+18] (e.g., using positron emission tomography or magnetic resonance imaging) on the full brain scans of predominantly healthy patients as part of a preventive healthcare program of a hospital. As always, it is natural to assume the availability of relevant medical information for each patient. Even though for most of the patients these images would depict a healthy brain function, it would be of great value to be able to ask an algorithm to describe sub-populations with potentially abnormal structures in their brain, which potentially indicate a pathological brain function.

Query Q3. *Which patient profile has the most abnormal brain scans?*

An answer to such a question can be of great value in exploring potentially new underlying causes of brain function abnormalities [DP16], potentially even as of yet unidentified. We can also consider the frequent cases where the patient imaging data is extended with other data modalities, such as transcriptomes; then, we can solely use the dissimilarity of the imaging data as an indicator of brain abnormalities, which, through the additional data modalities can be translated into novel associations between genes and abnormal brain function [AKC+21]. That comes in contrast to the typical procedure. In this, one would first have to i) focus on a specific hypothesis for a potentially affected brain function of interest, then ii) capture this into a scalar score that measures this abnormality for each patient, and only then iii) study transcriptomic associations with the derived brain function metric [ZSY+21]. We address this need for structure within the studied entities by proposing a novel framework for subgroup discovery where we lift the constraint of the scalar target; instead, we allow each entity to be associated with a target variable with virtually any type of a domain, as long as a suitable similarity function over this domain is defined.

Returning once more to our attempt to answer Query Q3, a second limitation of typical subgroup discovery becomes apparent. For each specific application of the existing works in typical subgroup discovery, the expert and the researcher has to perform several steps to design a practical method:

i) for each entity, a target variable needs to be specified,
ii) a distribution for the target must be assumed,
iii) a measure is needed to quantify the distribution distance,
iv) an appropriate optimisation algorithm for the measure needs to be instantiated,
v) which entails a tedious and cumbersome process to develop an efficient optimistic estimator for its practical exact optimisation.

In this work we break with this standard recipe, and instead propose a single algorithm that obviates the need for all these steps, while it also seamlessly allows for the incorporation of arbitrarily structured entities. To do so, we show how to employ a positive definite kernel defined over the structure associated with the entity. This allows for a profound paradigm shift, where we do not anymore have to devise a specialised objective function for some scalar score of the entities. Instead, we propose both an objective function that can measure the distance between the distributions of two entity samples, and an efficient optimistic estimator for this objective function, both of which depend solely on a positive definite similarity matrix between the entities. Thus, we can directly apply subgroup discovery on any population of entities with an arbitrary structure, as long as there is any positive definite kernel on this structure. Now, this simply involves plugging in our method the resulting Gramian of the chosen kernel. In other words, our research goal can be described as follows.

Goal G3. *Find the subgroup whose entities are deemed the least similar to those in the rest of the dataset, using an arbitrary positive definite kernel defined over entities.*

Importantly, this formulation allows us to also recover several cases of typical subgroup discovery as a special case within our framework, simply by using the linear kernel on the scalar target variable; in fact, this also yields the same computational complexity as in the typical case. Another family of kernels that, as such, can be used in our framework, can be derived by first generating an embedding of the entities in Euclidean space, and then applying an appropriate kernel on this embedding space, such as the simple linear kernel which corresponds to the standard Euclidean distance [PV11]. Importantly, the use of pre-computed embeddings also allows to seamlessly incorporate prior knowledge of each specific domain, as computed by unsupervised manifold learning methods such as [HTF09] principal

components analysis, multi-dimensional scaling, locally-linear embeddings, and t-stochastic neighbour embedding (t-SNE) [vdMH08]; also included are methods for structured data that depend on neural architectures, such as word2vec [MCC+13], node2vec, graph2vec, X2vec [Gro20].

Of course, the generality of our approach does not come for free, but requires the choice of an appropriate kernel. What further increases the difficulty of the problem, is that in this novel task there are no clear cut class labels that can be used for standard hyper-parameter optimisation; this means that standard metrics, such as the class accuracy, cannot be applied to select an appropriate kernel for this task. For this reason, we complete our method by proposing a novel measure of kernel fitness that takes into account the important parts of the available attribute information. Our measure can thus replace the classification accuracy in a standard hyper-parameter search, such as Bayesian cross validation. Importantly, our measure is also differentiable and allows closed form solutions for the multiple-kernel learning setting.

We present our framework of this modular similarity-based objective in Chapter 5, and we also present an algorithm to solve it for any choice of a positive definite kernel. What is more, we also discuss hyper-parameter optimisation methods to choose a kernel for each dataset, which includes a multiple kernel learning scenario.

1.3.4 A Flexible Random Walk Kernel

This above contribution allows us to utilise any of the extensive arsenal of kernels on structured data, which makes subgroup discovery applicable out-of-the-box on virtually any type of structured entities. Out of these, we lay a special interest on graphs, due to their ability to describe a variety of entities, and particularly in emerging fields such as molecules in computer aided drug discovery [GDD06], protein-protein interactions in biology [YFS+20], and in several omics analyses [LCC+21].

One of the established kernels on graphs is the family of random walk kernels [VSK+10], which compare two graphs based on a notion of similarity between random walks that are simultaneously performed over these two graphs. Motivated by the needs of our work, we further study an unexplored set of instances of this family, for the case when integer labels are available for each node, like those arising from brain connectome analysis, in which each graph node represents a region, whose function is similar to regions with nearby numerical index.

This property, of integer vertex labels that indicate similar vertex structure, can also arise naturally from other features in undirected graphs; our claim is

hence that simultaneous walks should take into consideration this similarity and only be performed on graph vertices that are closely related. This gives rise to the **S**tr**u**ctural **S**imil**a**rity ra**n**dom walk (SUSAN) graph kernel that we propose in Chapter 6. The key idea of SUSAN is that it takes into consideration this relationship of each pair of vertices from the two graphs during the simultaneous walks, and demonstrates its membership in the same framework of the random walk families by an appropriate re-formulation of this intuition. In this way, SUSAN fills the gap between the two extremes of full connectivity and only identical connections, both cases previously studied in the literature.

As an additional contribution, we propose ways to efficiently compute this instance of random walks, which for sparse, low rank graphs can become orders of magnitude faster than the naïve alternative. Importantly, when the distance of the integer labels becomes zero after a threshold, we additionally show that a certain block-banded matrix structure arises in the key computation of this kernel, and we provide a highly efficient C++ implementation of its vectorised computation, that can be orders of magnitude faster than state-of-the-art efficient implementations of the same computation, in which this banded structure is not taken into account.

Since SUSAN is a graph kernel, its applicability is by no means limited within subgroup discovery, but instead can be used in any machine learning application on graphs. Therefore, we also study its classification performance compared to a state of the art graph kernel [TGL+19] when used within a support vector machine, and we show the regimes where its accuracy is significantly superior in the statistical sense.

1.4 Contributions

A large part of the contributions of this work consists of novel methods that extend the domain of applicability of subgroup discovery into the large class of data with structured target concepts, which lies beyond what is currently possible using existing methods. At the same time, these contributions are not limited within subgroup discovery. These contributions are as follows.

Contribution 1. *We perform an in-depth analysis of the theoretical interpretation of the various objective functions used in typical subgroup discovery, which includes the established impact function and its special cases. More specifically, we provide a unified landscape in which we position existing methods, including our own contributions, and make a link to the statistical implications of common decisions in existing subgroup discovery methods, even when this analysis is missing in the original works.*

Contribution 2. *We propose a novel algorithm for the efficient and exact optimisation of our problem and compare it to existing alternatives, while we also provide a mathematical description of the optimisation domain. Importantly, we re-interpret an established method from a multi-objective optimisation approach, in which we provide a mathematical explanation of the involved parameters—previously left up to the user intuition—and show the structure of the resulting optima as a particular subset of the Pareto frontier between key objectives. We use this approach to present all results in this work.*

Contribution 3. *We propose the concept of representative subgroups, and thoroughly study its application on generating descriptions that correspond to a user-specified extent to different classes of an additional variable, that captures special attributes of the data. To this end we also propose **R**epresentativeness **Aw**are algo**r**ithm (RAWR), a novel algorithm for the optimal and efficient discovery of representative subgroups. We showcase the implications and usefulness of our method both as a means to overcome known trends, but also to ensure fairness in the resulting subgroups.*

Contribution 4. *We study the limitations of the established degree-based density measures of subgraph density, easily demonstrated through a realistic counter example. We correct these limitations by proposing a novel measure of density: robust connectedness, which incorporates more structural information per edge than the former measures, and provide an intuitive interpretation for its value. We develop **Ro**bustly–Connected **S**ubgraphs with Descr**i**ptions (RoSi), an exact and efficient algorithm that finds the subgroup of entities which form the most robustly connected subgraph.*

Contribution 5. *We introduce the use of positive definite kernels to measure the similarity between the studied entities, which constitutes the first kernel method for subgroup discovery. This approach allows us to extend the framework of subgroup discovery to virtually any kind of entities that go beyond simple scalars. For use in this task we propose an intuitive and novel objective function that is motivated by a well established statistic for the two-sample problem. We additionally develop techniques for the hyper-parameter tuning of necessary kernels, to enable the practical application of this method in real world settings.*

Contribution 6. *We propose a novel graph kernel for use in cases where alignment information is available for the graph vertices, which formalises the intuition of allowing a tuneable amount of leeway in the alignment of dissimilar vertices, based on the similarity of their labels. We additionally*

provide an efficient method to extract meaningful structure-aware vertex labels which allow for a distance-based comparison between the labelled vertices, and can be directly used by our kernel instead of alignment information. We show that the resulting graph kernel is a special case of the positive definite random walk kernels. We also propose efficient ways to compute this kernel.

Contribution 7. *We make openly available all source code of our methods, that enable their applications on practical settings. Importantly, we publish* SERGIO*, an interactive package for subgroup discovery that makes available most of our methods out-of-the-box, while remaining an extensible framework for the research and development of novel methods. For the needs of this work and as demonstrative cases, we also compile three datasets containing stocks listed in the NYSE (`Stock`), drug-like chemicals (`Chem`), and a sample of the twitter social media (`Twitter`), while we curate several others from publicly available repositories.*

Most of these contributions have appeared in the following conference publications; an analytical listing is available in Table 1.1.

1. Janis Kalofolias et al. Efficiently Discovering Locally Exceptional Yet Globally Representative Subgroups. In *2017 IEEE International Conference on Data Mining (ICDM)*, November 2017

2. Janis Kalofolias et al. Discovering robustly connected subgraphs with simple descriptions. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1150–1155, November 2019

3. Janis Kalofolias et al. SUSAN: The structural similarity random walk kernel. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 298–306. Society for Industrial and Applied Mathematics, January 2021

4. Janis Kalofolias and Jilles Vreeken. Naming the most anomalous clusters in Hilbert space for structures with attribute information. In *AAAI National Conference of the American Association for Artificial Intelligence*, 2022

The author of this dissertation is the first author in all these publications and has been responsible for the ideation of the posed problems and the development of the solutions to them; this included providing the theoretical proofs that support the correctness of the proposed methods, the implementation of the proposed algorithms and the orchestration of the appropriate experiments, and finally taking the lead in writing the corresponding manuscripts.

Public.	Contr.	Manuscript	Brief Description
[KBV17]	2	Section 2.4.2	• IDDFS algorithm • Its extensions
[KBV17]	3	Chapter 3	• Representative subgroups • RAWR exact algorithm
[KBV19]	4	Chapter 4	• Robust connectedness • RoSi exact algorithm
[KV22]	5	Chapter 5	• Kernelised subgroup discovery • Kernel fitness measure • Multiple kernel learning methods
[KWV21]	6	Chapter 6	• SUSAN graph kernel • Efficient iterative algorithms • Prototypical C++ implementation
[KV22]	7	Appendix B	• Novel datasets: `Stock`, `Twitter`, and `Chem` • SERGIO framework

Table 1.1: Index of the contributions within the published works of the author.

1.5 Outline

This manuscript introduces with the desired traits of the methods that discover interesting subgroups from a given dataset.

Next, in Chapter 2 these ideas are formalised into the framework of subgroup discovery, its key features, common objective functions and our contributions on optimisation algorithms, both for single and multiple objective optimisation.

Following this, in Chapter 3 we introduce the concept of representative subgroup discovery, we motivate its necessity and provide a framework to find the subgroup which, while being exceptional, retains a representative control distribution.

In Chapter 4 we continue with the introduction of a novel measure of robust connectedness, which is superior to the typically used density measure—we then present our RAWR method that finds the subgroup whose entities are most robustly connected.

Subsequently, in Chapter 5 we introduce the first method for kernelised subgroup discovery, which constitutes a versatile framework that only requires a pre-computed Gramian; we additionally provide a measure for the hyperparameter tuning of the involved kernel and solutions for related multiple kernel learning.

We continue in Chapter 6 to present in depth the SUSAN graph kernel, a member of the broad random walk kernels that allows for a flexible trade-off between strict and looser alignments between nodes of integer-valued labels; we also propose a novel scheme for deriving such labels in case of their absence.

Finally, we round up in Chapter 7 with a summary of this work and reflect critically on each proposed method and contribution. We conclude with an outlook into the future, including an outline of potential future work.

We note that for readability and conciseness we postpone any lengthy proofs to Appendix A. Additionally, in Appendix B we provide in-depth technical details, such as experimental settings, implementation details and elaborate on the details of the used datasets.

2 Subgroup Discovery

Subgroup discovery is a powerful framework of methods that aim to reveal the important local structure within a given dataset. Within this framework, we focus in particular on the data-centric[1] methods, whose goal is to find specific sub-parts within a dataset that are maximally deviating when compared to the behaviour of the rest (or the entirety) of the dataset. In this chapter we work toward a formalisation of this task and present the general framework for the exact optimisation of the associated objectives.

The datasets we study in this setting describe a population (or sample thereof) consisting of n entities E, for each of which we are also given a set of meaningful attributes $\mathcal{X}$. In the previously discussed healthcare application, our dataset was a sample of hospitalised patients and exemplary meaningful attributes were tracking whether the patient is a smoker, their sex, age, etc. These attributes are used to select **subgroups** from the population[2] E: sub-populations that are both exceptional and have an intuitive name. Formally, these attributes can be seen as functions $l_\rho : E \to \mathcal{V}_\rho$, that is, to every entity $\epsilon \in E$ each attribute $l_\rho \in \mathcal{X}$ assigns the value $l_\rho(\epsilon) \in \mathcal{V}_\rho$.

2.1 Deriving Predicates

We describe population subsets by using the attributes $\mathcal{X}$ as follows. First, the domain of each attribute is partitioned appropriately, and from each partition we derive meaningful predicates; for instance from the attribute "*smoker*" we can derive predicates "*smokes*", "*does not smoke*", while from the attribute "*age*" we could derive ones like "*under* 18", "*under* 40", "*not under* 18", etc. In this way we assemble a collection of m meaningful predicates

[1] We recall from Section 1.2 that the data-centric methods are the counterpart of the user-centric ones, which can be seen as a—typically heuristic—extension of the former ones. Although this formal introduction still applies to the user-centric methods, these methods typically lack desired properties, such as statistical interpretations and exact optimisation algorithms, and therefore remain outside the scope of our interest.

[2] Here, for simplicity we assume that the dataset contains the entire population. However, even when the dataset contains a sample of the population, as long as it is a uniformly sample populous one, we can still reasonably assume that any description that selects a sample of the dataset corresponds to a similar subset of an entire population—thereby a sub-population. We explicitly address this detail in Section 2.3, below.

P. Formally, each predicate p is a Boolean function $p_j : E \to \{\top, \bot\}$, where $\top$ and $\bot$ are the *true* and *false* conditions, respectively. More often than not, the focus will be on the name of the predicate, in which case we use an Iverson-like notation[3] $[\epsilon \text{ satisfies } \times] \equiv p_\times(\epsilon)$, often even dropping the mention of the entity ϵ; for instance, we hence represent the exemplary predicates as $[\texttt{smokes}]$, $\neg[\texttt{smokes}]$, $[\texttt{age} \leq 18]$, $[\texttt{age} \leq 40]$, and $\neg[\texttt{age} \leq 18]$.

Note that in this language we do not explicitly include disjunctions, since they do not only substantially complicate the optimisation procedure, but primarily also quickly deteriorate the intelligibility of the resulting descriptions. By forming the conjunction of negated predicates, however, we can still express a subset of disjunctive statements. Note that when this procedure is applied to predicates of the same attribute, this includes the case of **internal disjunctions** [Ray99]. These can be seen as special predicates added to the language, whose value is equal to the disjunction of several sub-predicates, here of the same attribute, e.g., $[13 \leq \texttt{age} \leq 18] \vee [\texttt{age} \geq 60] \equiv \neg[\texttt{age} < 13] \wedge \neg[18 < \texttt{age} < 60]$.

The derivation of predicates P is essentially equivalent to the discretisation of the domain of each attribute, and different schemes can be employed depending on this domain. Boolean attributes give rise to two predicates: one that is equal to the attribute itself and another equal to its negation; for categorical attributes a natural choice is to assign one predicate per category, e.g. for a virus strain we could derive $[\texttt{strain} = A]$, $[\texttt{strain} = B]$, and so on. For numerical attributes typically unsupervised methods are used and as a pre-processing stage [MK21]; among them, perhaps the most popular is **equi-quantile** discretisation, which uses bins of the same frequency. This method is applicable both as a parametric method—by first fitting the assumed distribution—or as a non-parametric one, which is equivalent to empirical histogram equalisation. An equally popular method is **equi-distant** discretisation, which can be seen as the special case of parametric equi-quantile discretisation for uniform distributions. Another notable mention is the Lloyd-Max [WKR+08] algorithm that computes the bins that minimise the mean-squared error, and is equivalent to applying k-means clustering for uni-variate discretisation. Although from the discretisation perspective subgroup discovery can be seen as a supervised method, the use of supervised discretisation has not been widely studied. Similarly limited works have also suggested on-the-fly discretisation schemes [GR09; MK21], out of which one sacrifices the optimality and is non-relevant [MK21], while the other

[3] The Iverson bracket [Ive62] is a compact representation of the Kronecker delta. Here, the bracket retains its property of a predicate, which allows for a convenient way to concisely represent descriptions of predicates and their conjunctions as logical expressions.

comprises a straightforward extension[4] to all our algorithms, but is beyond the scope of this thesis [GR09]. In this work we adopt the non-parametric variant of equi-quantile discretisation due to its versatility, where we typically use 5 intervals; this allows the resulting ranges to be named intuitively as *very_low*, *low*, *normal*, *high*, and *very_high*.

2.2 The Language of Subgroups

The predicates in P can be combined to form finer descriptions for the candidate sub-populations; the set of all such possible descriptions comprise the **subgroup language**. We now describe this set, for which we borrow useful notions from the field of formal concept analysis.

As Boolean functions that predicates are, each of them can be combined to form conjunctions, thus yielding a more fine-grained Boolean functions, and thereby predicate, that describes the membership of each entity in a sub-population. For instance, whereas $[\texttt{age} \geq 18]$ describes adults and $[\texttt{sex} = \textit{female}]$ females, when combined to form $[\texttt{age} \geq 18] \wedge [\texttt{sex} = \textit{female}]$ they yield the finer description "*adult women*". Formally, these Boolean conjunctions can also serve as indicator functions, which then "select" those entities from the sub-population that satisfy the Boolean condition or, equivalently, that fit the corresponding description. We therefore refer to these conjunctions as **selectors** and represent them as predicate subsets, always implying that there is an associated description that can be easily read from this predicate subset.

Each selector $s \subseteq P$ induces a sub-population through the **extension** operator $\text{ext} : 2^P \to 2^E$, where

$$\text{ext}(s; E) \coloneqq \{\, \epsilon \in E \mid p(\epsilon) = \top \text{ for all } p \in s \,\}\,, \qquad s \subseteq P\,. \tag{2.1}$$

An operator to a direction opposite to that of the extension, albeit not necessarily its inverse, is the **intention** operator $\text{int} : 2^E \to 2^P$. This operator assigns to every entity subset Q those predicates that are satisfied by all entities of Q

$$\text{int}(Q; P) \coloneqq \{\, p \in P \mid p(\epsilon) = \top \text{ for all } \epsilon \in Q \,\}\,, \qquad Q \subseteq E\,. \tag{2.2}$$

A key property of these two operations is that, once we append one or more

[4] This method keeps a 2-dimensional table of all combinations of upper and lower bounds for each attribute, which are combined to prune non-promising intervals. Since the methods we propose in this thesis provide algorithms that compute similar bounds, this scheme is also applicable to our tasks.

predicates to a selector s, the new extension is a subset of the original; similarly, when we append one or more entities to an entity subset, the new intention is a subset of the original. This property is formally expressed by saying that the two operators of extension and intention form an (antitone) Galois connection [GW99] between the powersets of entities and predicates

$$Q \subseteq \mathrm{ext}(s; E) \iff s \subseteq \mathrm{int}(Q)\,. \tag{2.3}$$

This is the property that we will later use to derive an algorithm that can efficiently optimise over the studied subgroup languages and is will also be the basis of a key component in the definition of a redundancy-free subset of the established subgroup language.

We can now define useful variants of the **subgroup language** $\mathcal{L}$. A first attempt would be to simply include every possible predicate conjunction; using the representation of conjunctions as predicate subsets, this yields the definition $\mathcal{L}_{\mathrm{all}}(P) \coloneqq 2^P$. In certain attribute configurations, however, it can happen that the same sub-population accepts several descriptions. This can happen when an added predicate does not further restrict the sub-population, e.g., $[\texttt{pregnant}] \equiv [\texttt{pregnant}] \wedge [\texttt{sex} = \textit{female}]$, where being pregnant already implies the sex. The same can also happen in more complex scenaria, when one description implies the other; assuming, for instance, that one can only obtain immunity from a disease by either a vaccination or naturally recovering from a historical exposure; we can now that we get the two equivalent descriptions $[\texttt{exposed}] \wedge [\texttt{healthy}] \equiv \neg[\texttt{vaccinated}] \wedge [\texttt{immune}]$. This multiplicity of descriptions creates a problematic redundancy that may lead to reporting the same result multiple times, while it additionally imposes a substantial yet avoidable overhead during optimisation.

One way to avoid this redundancy is to unite all equivalent descriptions of $\mathcal{L}_{\mathrm{all}}$ into an equivalence class and form a new language with only one representative description from each equivalence class, thus forming a subgroup language with no equivalent descriptions. Here we present a rigorous adaptation of the redundancy-free language defined by Boley and Grosskreutz [BG09]. Assuming we are given a sufficiently representative population E we can define descriptions as equivalent $\equiv_E$ when they have the same extension in this population

$$s_1 \equiv_E s_2 \iff \mathrm{ext}(s_1; E) = \mathrm{ext}(s_2; E)\,, \qquad s_1, s_2 \subseteq P\,. \tag{2.4}$$

Thus, the equivalence relation $\equiv_E$ partitions the members of $\mathcal{L}_{\mathrm{all}}$ forming

the quotient set[5]

$$\mathcal{L}_{\text{all}}/\equiv_E \coloneqq \left\{\{s \in \mathcal{L}_{\text{all}} \mid s' \equiv_E s\} \mid s' \in \mathcal{L}_{\text{all}}\right\}, \tag{2.5}$$

each member of which is a partition of $\mathcal{L}_{\text{all}}$ containing equivalent selectors. Now, any set of selectors that contains at most one member from each set in $\mathcal{L}_{\text{all}}/\equiv_E$ can have no equivalent selectors, and is thus guaranteed to be redundancy-free. We can therefore form the redundancy free language $\mathcal{L}_{\text{cl}}$ by allowing exactly one representative selector from each equivalence class of $\mathcal{L}_{\text{all}}/\equiv_E$, which we achieve as follows. We revisit the Galois connection property and apply the intention operator on both sides of the first equation in Eq. (2.3), to observe that

$$\text{int}(\text{ext}(s; E); P) \subseteq s, \qquad \text{for all } s \subseteq P. \tag{2.6}$$

This lets us use the (inclusion-wise) maximal element in each equivalence class as the class representative. We define the **closure operator**

$$\text{clos}(s) \coloneqq \text{int}(\text{ext}(s; E); P) = \bigwedge \{p \in P \mid \text{ext}(p; E) \supseteq \text{ext}(s; E)\} \tag{2.7}$$

and call its fixed points the **closed** selectors, which are unique within each equivalence class. We can now define as a redundancy-free subgroup language the one consisting of the closed selectors

$$\mathcal{L}_{\text{cl}}(P, E) \coloneqq \{s \subseteq P \mid \text{clos}(s) = s\}. \tag{2.8}$$

Despite solving the problem of redundancy, however, this language contains selectors that are not necessarily good candidates to form a description for human audiences. More specifically, we notice from the definition of the closure operator in Eq. (2.7) that every selector $s \in \mathcal{L}_{\text{cl}}$ contains all the predicates that do not constrain its extension $\text{ext}(s)$, which can be prohibitively many. In contrast, we would like to be able to describe the very same extension but using only a few predicates, thus yielding a succinct

[5]Here we assume all equal elements in this formulation to be included only once. A more rigorous albeit overly complicated definition that does not repeat any equivalence class can be given in iterative form for any set in the standard Zermelo-Fraenkel theory with the axiom of choice (ZFC). We let $\mathcal{L}_0 \coloneqq \{s \in \mathcal{L}_{\text{all}} \mid s \equiv_E \min \mathcal{L}_{\text{all}}\}$, $\mathcal{L}_i \coloneqq \left\{s \in \mathcal{L}_{\text{all}} \mid s \equiv_E \min\left(\mathcal{L}_{\text{all}} \setminus \bigcup_{j=0}^{i-1} \mathcal{L}_j\right)\right\}$ and $\mathcal{L}_{\text{all}}/\equiv_E \coloneqq \bigcup_{i=0}^{I}\{\mathcal{L}_i\}$, where $\min \cdot$ is some arbitrary element of the set $\cdot$, and always exists by invoking the axiom of choice. Our case is much simpler, since such an element can be the minimal in terms of lexicographic ordering of the predicates contained in each selector, which is a well-defined total ordering for $\mathcal{L}_{\text{all}}$. Here, since $\mathcal{L}_{\text{all}}$ is finite and all $\mathcal{L}_i \neq \emptyset$, it is also $I < |\mathcal{L}_{\text{all}}|$.

description. In other words, by construction $\mathcal{L}_{\text{cl}}$ contains the inclusion-wise representative from each equivalence class, and an improved language could be derived by replaced this representative with a more succinct one from each equivalence class. This gives rise to the notion of a **minimal generator**, that is, a selector of which no predicate can be removed without changing its extension. Moreover, all minimal selectors that have the fewer possible number of predicates are called **minimum generators**. Note that not only the minimal but also the minimum generators can be exponentially many for given equivalence classes, and therefore we cannot retain the redundancy free property of a language by simply allowing all minimal or all minimum generators in one language.

One solution would be to use $\mathcal{L}_{\text{cl}}$ and from each closed selector derive a minimal one, which can be easily done in polynomial time. The case for minimum generators, however, is much more difficult, since the problem of computing a **minimum generator** of an equivalence class can be shown to be NP-hard [BG09] by a Karp reduction of the minimum set cover problem. For instance, although in our previous example it is easy to go from both $[\texttt{exposed}] \wedge [\texttt{healthy}]$ and $\neg[\texttt{vaccinated}] \wedge [\texttt{immune}]$ to their closure, which contains all four involved predicates at the same time, it is very hard to find one of the former, succinct descriptions starting from the closure itself. Luckily, we will later show that finding a minimum generator can indeed be performed with minimal additional overhead during the traversal of the optimisation algorithm we propose in this work, by keeping track of the path from which we encountered each equivalence class of $\mathcal{L}_{\text{all}}$. This would be equivalent to using a subgroup language $\mathcal{L}_{\text{mg}}$ that is a particular instance within the collection of all possible redundancy-free languages that consist only of a single minimal generator from equivalence class.

Beside the obvious implications of choosing a subgroup language on the expressivity of the resulting descriptions, this language also defines the domain over which we seek the optimal description, and thereby the optimal subgroup. The next step is to formalise our intuition into an objective function that dictates which is the best subgroup within this domain.

2.3 Measuring Subgroup Quality

The goal of subgroup discovery is to find the subgroup which maximises a quality measure; we refer to any such measure as the **objective function**. In this work we focus on the data-centric methods within this framework, in which the property used to assess the fitness of candidate subgroups is how exceptional any given subset of the population is. Thus, the goal of the

objective function is to assess the exceptionality of each subset of entities[6] with respect to a **target concept**, that is, a property of interest defined on this subset of entities. Typically, the target concept is simply a scalar variable defined on each entity, hence referred to as the **target variable**; in this case, the exceptionality of the entity subset is assessed based on the statistics of this variable within the evaluated subset. Formally, for the target variable we write $y\colon E \to \mathcal{V}_y$, and its domain $\mathcal{V}_y$ can be either discrete or numerical. In the simplest—and most frequent—scenarios, the value of the target variable is simply provided in the dataset, similar to the available attributes. For instance, in our medical example, meaningful target variables would indicate whether the patient survived $y_{\text{surv}} \in \{\top, \bot\}$ or the duration of their hospitalisation $y_{\text{hosp}} \in \{0, 1, \dots\}$.

Using this target variable, we can define an important objective function, the weighted relative accuracy (WRAcc) [TFL00], that was first proposed for binary target variables, and has also been used with numerical targets under the name of **impact function** [Web01]

$$f_{\text{imp}}(Q) \coloneqq |Q| \cdot (\bar{y}_Q - \bar{y}_E)\,. \tag{2.9}$$

Here, $\bar{y}_Q \coloneqq \text{mean}\{y(\epsilon) : \epsilon \in E\}$ is the mean of the target values of all entities in the population E, $\bar{y}_Q$ is the target mean of only those entities in the candidate subset Q. The first intuition proposed to motivate this objective was that it combines two concepts that in our work we will broadly refer to as **generality** and **exceptionality**. More specifically, the f_{imp} objective favours larger subsets that are therefore less likely to be outliers (*generality*), while at the same time preferring those with greater mean deviation (*exceptionality*), which in turn makes them more outstanding. The two factors are multiplied to preserve the scaling of the objective with respect to each component, i.e., doubling any factor while keeping the other constant has the same effect on the final objective. This observation also reveals the relation to the titular notion of impact in physics, whose result remains the same whether we double the mass or the speed (keeping the other term constant), and where the mass and speed serve as parallels for the size and deviation, respectively.

In this work we will extensively study a generalisation of the impact function of Eq. (2.9): the geometrically weighted impact (GWI)

$$f_{\text{gwi}}(Q; \gamma) \coloneqq |Q|^{\gamma} \cdot \left([\bar{y}_Q - \bar{y}_E]_+\right)^{1-\gamma}, \qquad \gamma \in [0, 1]\,, \tag{2.10}$$

[6] Note that the objective function is not defined exclusively on the domain of subgroups, but can be applied on any subset of entities.

which introduces a parameter γ that tunes the geometric weight between the two terms of generality and exceptionality, and only uses the positive part[7] $[\cdot]_+ := \max(\cdot, 0)$ of the latter. This formulation already contains several proposed objective functions as special cases of its tuning parameter. For instance, for $\gamma = 0$ it becomes equivalent (in terms of maximisation) to the lift and the relative gain [Atz15], for $\gamma = 1/3$ it gives the standardised mean z-score [TLT08], while for $\gamma = 1/2$ we recover the original WRAcc and impact function. Due to the central position of this function in our work, we invest a portion of this analysis to motivate it, study its properties and demonstrate its importance.

A more principled approach to derive objective functions comes from the statistical interpretation of the task at hand. For this, we assume that the target variable corresponds to realisations of the random variable $\mathcal{Y}$, from which point onward we can derive sensible statistics adapted to common problems in statistical theory. To this end we denote $\mathbb{P}_{\mathcal{Y}}$ the distribution of the target attribute in the actual population and $y(Q)$ the set of target variables of entities within Q, which is considered a sample of the random variable $\mathcal{Y}_Q \sim \mathbb{P}_{\mathcal{Y} \mid Q}$. At this point it becomes clear that the statistics of $\mathcal{Y}_Q$ should be estimated based on all entities of the subgroup Q. However, we still need to define which entities we should use to estimate the distribution of $\mathcal{Y}$, against which the former estimate is to be compared. In this thesis we contrast the two main approaches and explicitly reason about them using statistical assumptions, which improves on what is standard in the literature. So far, the distinction between the two approaches is mostly ignored [HCG+11; Hel16] and the corresponding choice is implicitly made based on what is convenient for the computation or the theoretical model [SKF+16], or simply let to the intuition of the user [Atz15], for instance as part of the 'art' of choosing the right objective function [LAP16]. We hence explicitly distinguish two cases, depending on our statistical assumptions for the population available in the dataset.

Case I: Entire Population Accessible On one hand, it might be reasonable to assume that we know the true distribution of the population. This could

[7] Using the positive part of the exceptionality term is solely used for the added theoretical convenience and does not hurt the generality of f_{gwi} in practical scenarios; in fact other works also adopt a similar formulation of the impact function itself [BGG+17]. The reason for this is that all practical domains over which we perform optimisation contain the entire population E as a candidate subset, which attains a zero impact function $f_{\text{imp}}(E) = f_{\text{gwi}}(E, \gamma) = 0$, for all $\gamma \in [0, 1)$. Therefore, no subset with a negative f_{imp} or f_{gwi} could ever be an optimiser, regardless of the existence of the positive part $[\cdot]_+$ in the formulation of each of these objective functions.

happen, for instance, if we assume that the dataset contains the entire population, in which case we can easily compute important parameters of the distribution exactly, such as its probability mass function and its moments. This might also be a valid assumption when the distribution of the target concept is known, for instance if we knew that the target variable $\mathcal{Y} \sim \mathcal{N}(\mu, \sigma^2)$ is normally distributed. Then, we may assume the dataset to be large enough to allow the empirical estimation of the distribution parameters with sufficiently high confidence, so that we can consider these parameter estimates as deterministic quantities. Thus, here $\mathbb{P}_{\mathcal{Y}}$ is the distribution of the target attribute in the entire dataset.

Problem 1 (Goodness-of-fit Problem)**.** *Does the sample $y(Q)$ come from the distribution $\mathbb{P}_{\mathcal{Y}}$? Equivalently, is $\mathbb{P}_{\mathcal{Y}} = \mathbb{P}_{\mathcal{Y} \mid Q}$, given the single sample $y(Q)$?*

As an example of how such an objective function can be used, assume that $\mathcal{Y}$ $\mathcal{N}(\bar{y}, \sigma^2,)$, where $\bar{y} = \bar{y}_E$ is assumed to be the true mean of this distribution. Now an appropriate statistic to find the statistic appropriate for this problem is the z-score of the empirical mean. Its absolute value for i.i.d. samples becomes

$$|z(Q)| \coloneqq \frac{|\hat{\bar{y}}_Q - \bar{y}|}{\sigma/\sqrt{m}} = \frac{1}{\sigma}\sqrt{m}|\hat{\bar{y}}_Q - \bar{y}| \propto f_{\text{gwi}}^{3/2}(Q; 1/3)\,. \tag{2.11}$$

In other words, optimising the GWI function of Eq. (2.10) for $\gamma = 1/3$ we get the subset that would yield the lowest p-value for the two-tailed test with a null hypothesis assuming a normally distributed target variable.

Still well within this regime, a different argumentation to derive a statistic for categorical target variables comes from the use of proper scoring rules as summaries for the subgroup statistics. To this end, Song et al. [SKF+16] propose to assess the goodness of fit with the Brie and log-loss proper scoring rules. These yield the corresponding divergences

$$d_{\text{Brier}}(Q) \coloneqq \sum_{k=1}^{K} (p_k - q_k)^2 \qquad d_{\text{Log-Loss}}(Q) \coloneqq \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k}\,, \tag{2.12}$$

where we assume the target variable to be categorical with K classes with p_k, q_k denoting the probabilities of class k in the subgroup and the true population, respectively

$$p_k \coloneqq \mathop{\mathbb{E}}_{y \sim \mathbb{P}_{\mathcal{Y} \mid Q}} \Big[\mathbb{1}[y = k]\Big] \quad q_k \coloneqq \mathop{\mathbb{E}}_{y \sim \mathbb{P}_{\mathcal{Y}}} \Big[\mathbb{1}[y = k]\Big] \quad k = 1, \ldots, K\,. \tag{2.13}$$

Their final proposed statistic is the information gain for each divergence, which gives the above divergences multiplied by the subgroup size m. Note that in the case of the Brier score, the resulting statistic is the multi-class analogue of the absolute value of the z-score.

Case II: Only a Population Sample Available Another approach for our task is to assume that the dataset is only a sample of the true population, arguably an assumption that needs not much reasoning to justify. Since then we do not know the true distribution of the population, we may only assume the dataset to have a sample of it, denoted $y(E) \sim \mathbb{P}_{\mathcal{Y}}$. This implies that to avoid any bias between the samples, we may not share instances between the two samples. That means that, since we should use all samples of Q to estimate a statistic for $\mathbb{P}_{\mathcal{Y}\,|\,Q}$, we may only use the complement $E \setminus Q$ to estimate any statistic for $\mathbb{P}_{\mathcal{Y}}$.

Problem 2 (Two sample Problem)**.** *Do the samples $y(Q)$ and $y(E \setminus Q)$ come from the same distribution $\mathbb{P}_{\mathcal{Y}}$? Equivalently, is $\mathbb{P}_{\mathcal{Y}} = \mathbb{P}_{\mathcal{Y}\,|\,Q}$, given independent samples $y(E \setminus Q) \sim \mathbb{P}_{\mathcal{Y}}$ and $y(Q) \sim \mathbb{P}_{\mathcal{Y}\,|\,Q}$?*

We will study a very versatile statistic for this latter problem in Chapter 5. Importantly, statistics developed for the two-sample problem can also be applied to address the goodness-of-fit problem, where we simply use as sample of the distribution $\mathbb{P}_{\mathcal{Y}}$ that from the entire population $y(E)$, albeit with the danger of possibly introducing bias in the resulting statistic.

Summarising the above, not only does the data-driven paradigm of subgroup discovery arise from well-founded statistical principles, but our adopted objective contains the centrally positioned (absolute of the) z-score.

2.4 Optimising the Objective Function

From this perspective we can recognise subgroup discovery as an optimisation problem, and in fact a hard combinatorial one. Formally, we can express this optimisation problem as

$$Q_{\text{opt}} \coloneqq \underset{Q \in \text{ext}(\mathcal{L};E)}{\arg\max}\ f(Q)\,, \tag{2.14}$$

where we overload notation to denote $\text{ext}(\mathcal{L};E) = \{\text{ext}(s;E) \mid s \in \mathcal{L}\}$ the set of all subsets of E corresponding to a subgroup in the chosen language. This is a constrained combinatorial problem, and as such its difficulty not only depends on the properties of the objective function f, but also on the structure of the optimisation domain $\text{ext}(\mathcal{L};E)$. We now present 3 algorithms

to solve it, in increasing order of versatility, out of which the last one is the one developed and used in this work.

2.4.1 Mathematical Programming

When the objective function allows it, we can formulate the problem of Eq. (2.14) as a standard mathematical program. This allows the use of very well studied algorithms available in highly optimised solvers that can be applied to the problem with little to no extra overhead.

For the sake of simplicity we focus this first exposure on the case of an integer linear program (ILP); later we will use the same principles to solve both a fractional program and a quadratic one. We therefore consider first the objective function of Eq. (2.9), from which we drop the absolute[8] value. To proceed we first assume an arbitrary total ordering $\leq$ of the elements in E, so that we can express any subset Q of $E = \{\epsilon_1, \ldots, \epsilon_n\}$ using its indicator vector

$$\mathbf{x}_Q = \mathrm{ind}_{\leq,E}(Q) \coloneqq \left(\mathbb{1}[\epsilon_i \in Q]\right)_{i=1}^{n} \qquad \mathbf{x}_Q \in \{0,1\}^n\,, \tag{2.15}$$

$$Q = \mathrm{set}_{\leq,E}(\mathbf{x}_Q) \coloneqq \bigcup_{\substack{1\leq i\leq n\\ \mathbf{x}_i=1}} \{\epsilon_i\}\,, \tag{2.16}$$

where $\mathrm{set}(\mathbf{x}_Q)$ is the set corresponding to the indicator vector $\mathbf{x}_Q$, and $\mathbb{1}[\cdot]$ is the indicator function that yields 1 if condition $\cdot$ is true and zero, otherwise. Using this notation we can now write $|Q| = \mathbf{e}^\top \mathbf{x}_Q$ and $\bar{y}_Q = \frac{\mathbf{e}\mathbf{y}}{\mathbf{e}^\top \mathbf{x}_Q}$, where $\mathbf{y} \coloneqq \left(y(\epsilon_i)\right)_{i=1}^{n}$ is the vector of all target values. Now our optimisation problem can be written as

$$\begin{aligned}\max_{Q\in\mathrm{ext}(\mathcal{L};E)} f_{\mathrm{imp}}(Q) = \\ \max_{\mathrm{set}(\mathbf{x}_Q)\in\mathrm{ext}(\mathcal{L};E)} \mathbf{e}^\top\mathbf{x}_Q \left(\frac{\mathbf{y}^\top\mathbf{x}_Q}{\mathbf{e}^\top\mathbf{x}_Q} - \bar{y}\right) = \\ \max_{\mathrm{set}(\mathbf{x}_Q)\in\mathrm{ext}(\mathcal{L};E)} {\mathbf{x}_Q}^\top(\mathbf{y} - \bar{y}\mathbf{e})\,,\end{aligned} \tag{2.17}$$

in which the objective is linear, since the vector in the parenthesis is independent of $\mathbf{x}_Q$. All that is left to express this problem in the standard form of an ILP is to provide a set of linear inequalities that define some polytope which contains all combination of $\mathbf{x}_Q$ that correspond to acceptable subsets

[8] Note that being able to optimise the objective problem without the absolute value is more powerful, as we can also obtain the optimum of the

$Q \in \mathrm{ext}(\mathcal{L}; E)$. For this we will use an additional vector of helper variables $\mathbf{p}_s \in \{0,1\}^m$ that will serve as the indicator vector of the predicates that are currently selected.

Lemma 2.1 (The language polytope)**.** *The indicator vectors of all sets of all named entity subsets that comprise the predicate language $\mathcal{L}_{\mathrm{all}}(E,P)$ are contained in the intersection of the integer lattice $\mathbb{Z}^{n+m}$ with a polytope $L_{E,P}$*

$$Q \in \mathrm{ext}(\mathcal{L}_{\mathrm{all}}(E,P); E) \iff (\exists \mathbf{p}_s) : (\mathbf{x}_Q, \mathbf{p}_s) \in \mathbb{Z}^{n+m} \cap L_{E,P}\,, \qquad (2.18)$$

where $\mathbf{p}_s \in \{0,1\}^m$ serves as a helper variable and corresponds to the indicator vector of a selector $s \subseteq P$, which in turn is a predicate subset.

The polytope $L_{E,P}$ itself is defined through the inequalities

$$-x_i - \sum_{j=1}^{m} p_j(1 - v_{i,j}) \leq -1 \qquad 1 \leq i \leq n \qquad (2.19)$$

$$x_i \sum_{j=1}^{m}(1 - v_{i,j}) \leq \sum_{j=1}^{m}(1 - x_i)(1 - v_{i,j}) \qquad 1 \leq i \leq n \qquad (2.20)$$

$$x_i \leq 1\,, \quad p_j \leq 1\,, \quad x_i \geq 0\,, \quad p_j \geq 0 \qquad 1 \leq i \leq n\,, \quad 1 \leq j \leq m\,, \qquad (2.21)$$

for $m = |P|$ the number of predicates and $v_{i,j} \coloneqq \mathbb{1}[p_j(\epsilon_i) = \top]$ the validity of the j-th predicate on the i-th entity. The optimal subset and corresponding selector can be read from $\mathbf{x}$ and $\mathbf{p}$, which serve as characteristic vectors over the set of entities and predicates, respectively.

Proof. Let $\mathbf{x}$ and $\mathbf{p}$ be the characteristic vectors of the entities and predicates, respectively. Now any subset $Q \subset E$ and any selector $\mathcal{L} \subset P$ can be described as points in $\{0,1\}^n \times \{0,1\}^m$.

We first show that the first two inequalities constrain the domain of the integer lattice exactly to the points corresponding to elements of $\mathcal{L}$. The first inequality ensures that if all selected predicates are validated by an entity, it must be selected. Indeed, when all selected predicates are valid for ϵ_i, the sum in the left hand side of the equality becomes 0, so the constraint is only valid when $\mathbf{x}_Q = 1$.
The second inequality enforces that if any predicate that is invalid for ϵ_i is selected, then x_i must be 0. These two conditions describe exactly all conjunctions of predicates, which coincides with the set $\mathcal{L}_{E,P}$. □

In practice, most software packages for integer programming allow the box constraints to be embedded in the specification of the variables, so that we

can limit the required inequalities to those of Eqs. (2.19) and (2.20), thus yielding the ILP

$$\begin{array}{ll} \max & \mathbf{x}^\top \mathbf{c} \\ \text{s.t.:} & A_L[\mathbf{x};\mathbf{p}] \leq b_L \\ & \mathbf{x} \in \{0,1\}^n\,,\ \mathbf{p} \in \{0,1\}^m \end{array} \tag{2.22}$$

with $\mathbf{c} = \mathbf{y} - \bar{y}\mathbf{e}$ and such $A_L \in \mathbb{R}^{2n \times n+m}$ and $b_L \in \mathbb{R}^{n+m}$ that encode the linear inequalities in Eqs. (2.19) and (2.20).

Despite its convenience, however, mathematical programming can not always be used for the objective functions desired, but can also be forbiddingly inefficient for more complex objectives, such as quadratic or fractional programs. This gives rise to the need for combinatorial algorithms that can take advantage of the special structure of the language polytope $L_{E,P}$.

2.4.2 An Efficient Branch-and-Bound Method

Since subgroup discovery is a combinatorial problem, a standard method for its optimisation is the branch and bound [MS08] algorithm. For the needs of our work we developed a dedicated branch and bound variant that we specialise for the optimisation over the domain $\mathrm{ext}(\mathcal{L};E)$. Below, we first describe the key components of branch and bound methods, and complete our specialisation in the last paragraph of this section.

The typical branch and bound approach efficiently traverses a spanning search tree over the elements of E, for which it combines two key components that handle the search node generation and the pruning, respectively.

Spanning Tree Generation The nodes of the search tree are dynamically generated by the **refinement operator** $\rho : \mathcal{L} \to 2^{\mathcal{L}}$, which for each subgroup in $\mathcal{L}$ computes a set of valid refinements, i.e., subsets of it that are also members of $\mathcal{L}$. In this way, the refinement operator induces a tree over $\mathcal{L}$—and thereby over $\mathrm{ext}(\mathcal{L};E)$—with the added requirement that this tree be spanning. In the following we assume some total ordering $\leq$ over P, e.g., the lexicographic ordering of their names, so that we can index the predicates as $p_1 \leq \ldots \leq p_m$.

When it comes to the subgroup language $\mathcal{L}_{\text{all}}$ of all predicate combinations, a simple refinement operator can be formulated by extending a given selector with each unused predicate. This gives a simple version of the **minimal description refinement operator**

$$\rho_{\text{md}}(s) \coloneqq \{s \wedge p_i \mid i_{\max}(s) < i \leq |P|\}\,, \quad i_{\max}(s) \coloneqq \max\{i \mid p_i \in s\}\,, \tag{2.23}$$

whose induced spanning tree reaches every partition of equivalent description multiple times, but always also from a minimal generator. Depending on the flavour of branch-and-bound in which it is employed, this can be a very useful property, when for example we seek to search over all conjunctions in $\mathcal{L}_{\text{all}}$ that are generated by no more than a fixed number of predicates.

Depending on the attribute configuration of the population, however, the redundancy in the search node generation can incur a considerable additional workload. In these cases, one can choose to traverse only the members of $\mathcal{L}_{\text{cl}}$. This can be achieved by the **redundancy-free refinement operator** [UAU+03]

$$\rho_{\text{rf}}(s) \coloneqq \left\{ s' = \text{clos}(s \wedge p_i) \mid s'|_{i-1} = s|_{i-1} \text{ for } i_{\text{core}}(s) < i \leq |P| \right\} , \quad (2.24)$$

which creates a spanning tree over exclusively the elements of $\mathcal{L}_{\text{cl}}$, thus avoiding any redundancy. In this tree, below each closed selector is a descendant with the same prefix p_j, and only if its extension is different than that of its parent. This is achieved by setting

$$s|_j = \bigwedge \{ p_j : p_j \text{ occurs in } s \text{ and } j \leq i \} \quad (2.25)$$

$$i_{\text{core}}(s) \coloneqq \min\{ j \mid \text{ext}(s|_j; E) = \text{ext}(s; E) \} . \quad (2.26)$$

Due to this prefix-preserving property, however, even though all equivalence classes of $\mathcal{L}_{\text{all}}/\equiv_E$ (see Eq. (2.5)) will eventually appear on this tree, it can be that a selector with a much more succinct equivalent will only appear in deeper layers of the tree. For this reason, this refinement operator cannot guarantee exhaustive search over the descriptions with limited number of predicates by limiting the depth of the induced search tree.

The induced spanning tree of both ρ_{md} and ρ_{rf} has as root the selector corresponding to the empty conjunction s_{root}, whose extension is the entire population E. In both cases, the algorithm traverses the spanning tree induced by the chosen refinement operator over $\mathcal{L}$, while pruning sub-optimal branches using an admissible bound of the objective function.

Tree Pruning Since the size of the search tree is exponential in the number of predicates, an efficient invocation of this optimisation algorithm requires an efficient scheme to prune the sub-optimal branches. To develop such a scheme we first observe that the refinement operator of a selector s appends to it additional predicates; due to the Galois connection property of Eq. (2.3), the extension of any refinement of s is a subset of $\text{ext}(s; E)$.

We can therefore bound the value of the objective function f over all nodes in the search sub-tree below selector s by the value of the **optimistic**

estimator $\hat{f}$. This is the term used in subgroup discovery for the upper bound of f over all subsets of the extension of the current selector $Q = \text{ext}(s; E)$,

$$\hat{f}(Q) \geq \max_{R \subseteq Q} f(R), \qquad \text{for all } Q \subseteq E\,. \tag{2.27}$$

Therefore, at each node of the search tree we can consult this bound and test if it is less than the objective value of the best subgroup found so far; if the bound is indeed less, then we can safely prune the entire search sub-tree below s, since we are guaranteed that no subset of Q and therefore also no subgroup in this sub-branch can improve on the best currently found subgroup. Naturally, as the bound of the optimistic estimator gets tighter its pruning potential increases. This potential becomes optimal when Eq. (2.27) holds with equality; then we refer to $\hat{f}$ as the **tight optimistic estimator** [GRW08] of the objective function f.

Since the value of the optimistic estimator is computed for every node in the search tree, its computation must be substantially more efficient than the original problem Eq. (2.14). This is much easier to achieve than solving the original constrained problem, due to both the unconstrained nature of the bound and the additional freedom to choose any upper bound, not necessarily the tight one. Nevertheless, there is no standard procedure for the derivation of an efficient optimistic estimator, and it requires a considerable effort for each corresponding objective function. In fact, this is one of the limitations of the exact methods for subgroup discovery, and one for which we provide a solution in Chapter 5.

Iterative Deepening Search All standard variants of branch-and-bound combine these components as follows. Starting from the empty selector s_{root} they traverse the search tree induced by the used refinement operator ρ, while keeping track of the best objective value of all selectors encountered so far. For each refinement of the current selector the optimistic estimator is evaluated, and if its value is below the current best the refinement is dropped, otherwise appended or pre-pended in the queue of active selectors. This description pertains to both breath- and depth-first searches, both of which have considerable downsides: During breadth-first-search the queue of active search nodes can grow arbitrarily large, which can readily incur a considerable memory overhead, while the depth-first variants can delve excessively deep in sub-optimal branches.

We therefore improve on these typical approaches, by introducing the iterative deepening depth-first search [Kor85] method, which is a variant of the classical branch-and-bound algorithm that uses a hybrid between depth-

Algorithm 1: Iterative deepening branch-and-bound .

Input: Result count k, depth limit $d_{\max}$, approx. factor α
Output: Top-k results $\mathcal{R}$

$\tau \leftarrow -\infty$, $\mathcal{R} \leftarrow \{\}$, $d_{\text{dfs}} \leftarrow 1$
do
  `truncated` $\leftarrow \bot$
  `stack` $\leftarrow$ **push**(**newStack**(), $(s_0, 0)$) // Initialise empty stack
  while notEmpty(`stack`) **do**
    `stack`, $(s_{\text{cur}}, d_{\text{cur}}) \leftarrow$ **pop**(`stack`)
    for $s_{\text{ref}} \in R(s_{\text{cur}})$ **do**
      $R \leftarrow \text{ext}(s_{\text{ref}}; E)$
      $\hat{f}_{\text{ref}} \leftarrow \hat{f}(R)$
      if $\hat{f}_{\text{ref}} > \alpha \cdot \tau$ **then**
        $f_{\text{ref}} \leftarrow f(R)$
        if $f_{\text{ref}} > \tau$ **then**
          $\mathcal{R} \leftarrow$ **keepTopK**$(\mathcal{R} \cup s_{\text{ref}})$
          $\tau \leftarrow \min\left\{ f\left(\text{ext}(s; E)\right) \middle| s \in \mathcal{R} \right\}$
        if $d_{\text{cur}} < d_{\text{dfs}}$ **then**
          push(`stack`, $(s, d_{\text{cur}} + 1)$)
        else
          `truncated` $\leftarrow \top$
  $d_{\text{dfs}} \leftarrow d_{\text{dfs}} + 1$
  return $\mathcal{R}$
while $d_{\text{dfs}} \leq d_{\max}$ **and** `truncated`

Truncated depth first search (DFS)

first and breadth-first searching. This variant iteratively invokes a truncated depth-first search with increasing maximal depth $d_{\max}$, so that shallow optimal solutions are found early. Since the inner invocation is essentially a depth-first search, it also requires only a minimal memory footprint of $O(|P|d_{\max})$ space, while having asymptotically the same complexity as both algorithms.

Optionally, and to achieve even better pruning, the branch-and-bound algorithm may use the relaxed comparison $\alpha\hat{f}(Q) > f(Q_{\text{opt}})$, for an approximation factor $\alpha \in (0, 1]$, where a value of $\alpha = 1$ yields the best subgroup. Lower α values generally result in more aggressive pruning, but only offer the relaxed guarantee that the discovered subgroup has a value no less than

α times that of the best subgroup. The general outline of this algorithm takes the form of Algorithm 1, which is the first contribution of this thesis.

More specifically, iterative deepening depth first search (IDDFS) starts with a permissive pruning threshold and empty result set (line: 1) and repeatedly invokes the inner truncated depth first search (DFS) (lines: 3-22). The latter traverses the tree induced over $\mathcal{L}$ by the chosen refinement operator ρ (line: 7) starting with the root selector s_{root} (line: 4). During traversal, the current selector s_{cur} is popped from the stack and for each of its refinements $s_{\text{ref}} \in \rho(s_{\text{cur}})$ we perform the following steps. First, we compute the optimistic estimate of the refinement (line: 9); if this estimate does not exceed the pruning threshold $\alpha\tau$, the refinement but also all its sub-refinements cannot provide an acceptable improvement over the current top-k and they are all ignored. Otherwise, if the objective value of the refinement improves on one of the current top k, we update this latter set as well as the threshold τ (lines: 13-14). Additionally, if the current depth allows, we append the current refinement to the stack (line: 16). This process repeats until the stack is empty, after which we restart the innermost DFS with an increased depth limit while maintaining the current pruning threshold τ. In this fashion, although consecutive DFS invocations still start from the same root selector s_{root}, as time progresses better results are found and τ increases, so that in subsequent invocations increasingly more nodes are pruned. This process repeats until DFS completes un-truncated, i.e., all reachable refinements have been traversed (line: 18).

From the study of the algorithm we see that the objective value and possibly the optimistic estimator must be computed once per iteration. Since the cost of each iteration just for the creation of the next studied refinement happens in (amortised) linear time, it is desirable for the bound to also be computable in close-to-linear time, in order to avoid changing the asymptotic complexity of the algorithm.

The branch-and-bound algorithm bears similarities with adapted versions of constraint satisfaction solvers, assuming the latter have been equipped with an appropriate constraint propagator that would take into account the same bound. The search tree of both algorithms can also be very similar. Their main difference lies in the way the search nodes are generated: in the constrained satisfaction approach a node is first created and then the constraints are applied, which often leads to the creation of infeasible solutions. In contrast, branch-and-bound uses node generation rules which guarantee the traversal of exclusively all subgroups in $\text{ext}(\mathcal{L}; E)$.

At this point we have presented ways to efficiently optimise a given objective function over a collection of useful subgroup languages. Importantly,

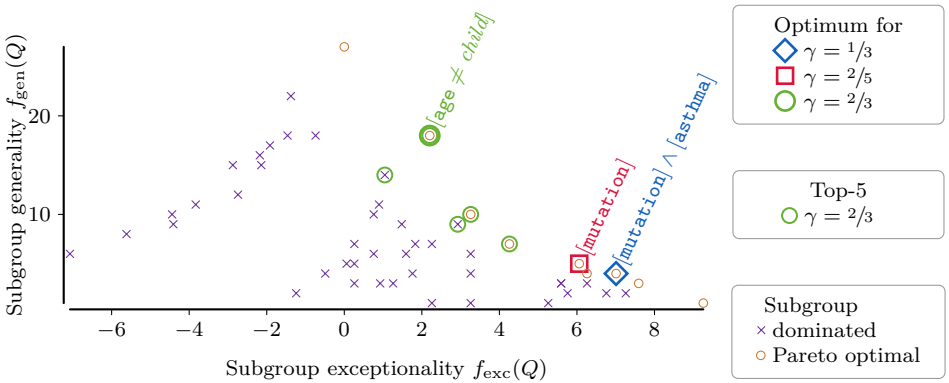


Figure 2.1 [Subgroup Objective Space]: The objective space of the 51 non-empty subgroups out of the total 243 that are defined by the attributes of the patient toy example (see Fig. 1.1a). Here we mark the subgroups of the Pareto frontier and the optimal subgroup of the GWI function for different relations between the two terms, as captured by this function for a varying γ parameter. Specifically for $\gamma = 2/3$ we additionally mark the top-5. We observe that changing the relation between the two terms selects a different point of the Pareto frontier: the upper-right "coast" of points. In contrast to the full Pareto frontier, which by definition spans the entire coast, the top-k points are redundantly close.

the methods of this section can also be generalised to go beyond the optimisation of a single objective function, into the optimisation of more than one objectives, as we present below.

2.5 Multi-objective Optimisation

All methods of data-centred subgroup discovery that we encountered so far have employed objective functions that assumed a fixed relation between a metric of subgroup generality and another of its exceptionality. However, despite their established usefulness[9], these methods yield a single optimal subgroup that may often be insufficient on its own, as it might be too general or too specific, or correspond to inaccurate or misleading statistical assumptions. The same also holds in the particular case of the geometrically weighted impact (GWI) function f_{gwi} in Eq. (2.10): even though this objective function allows the tuning of the relation between the two terms, its result

[9] We recall, for instance, that the GWI for a parameter of $\gamma = 1/3$ recovers the most significant subgroup when the target variables is assumed Gaussian, which can be of great use for medical personnel.

	Subset Description	f_{exc} $\bar{y}_Q - \bar{y}_E$	f_{gen} $\lvert Q\rvert$	Size %	Trade-off γ	Angle θ
1.	[age = *young*] ∧ [mutation] ∧ [asthma]	9.26	1	4%	[0.00 – 0.15]	2.1°
2.	[age ≠ *child*] ∧ [mutation] ∧ [asthma]	7.59	3	11%	[0.15 – 0.22]	7.7°
3.	[mutation] ∧ [asthma]	7.01	4	15%	[0.22 – 0.40]	11.1°
4.	[age ≠ *child*] ∧ [mutation]	6.26	4	15%	✗	12.4°
5.	[mutation]	6.06	5	19%	[0.40 – 0.44]	15.8°
6.	[age ≠ *child*] ∧ [asthma]	4.26	7	26%	✗	29.4°
7.	[asthma]	3.26	10	37%	✗	46.5°
8.	[age ≠ *child*]	2.20	18	67%	[0.44 – 0.94]	70.4°
9.	∅	0.00	27	100%	[0.94 – 1.00]	90.0°

Table 2.1 [Pareto Frontier of Subgroups]: A collection of subgroups that describe exceptional sub-populations in the toy dataset of Fig. 2.1 (marked with circles) for a varying emphasis on subgroup generality, in increasing order of this score. This list contains all Pareto optimal subgroups for the two scores for exceptionality and generality. Out of these subgroups, our method selects a more concise set of conspicuous Pareto optima; its members are indicated by the interval of the γ parameter for which the subgroup is optimal.

is nonetheless subject to the same limitations, since this relation is still controlled by a fixed, user-defined parameter.

To demonstrate these limitations, let us revisit the toy patient data of Fig. 1.1a, the attributes of which allow the creation of 243 subgroups[10]. To evaluate these subgroups we use the exceptionality and generality terms of the GWI function, denoted as $f_{\text{gen}}(Q) \coloneqq |Q|$ and $f_{\text{exc}}(Q) \coloneqq \bar{y}_Q - \bar{y}_E$, respectively. Using these metrics we can now depict the (non-empty) subgroups in the **objective space**, where each subgroup is represented as a point with coordinates $\big(f_{\text{exc}}(Q), f_{\text{gen}}(Q)\big)$ the values of the two terms, as shown in Fig. 2.1. Here, we can visually inspect the quality of the subgroups, which increases in terms of generality as we move to the top and in terms of exceptionality as we approach the right, with the optimum of each objective function determined by the relation between the two terms. For instance, when these terms are combined within the GWI function with a weight

[10] To see this, we note that there are 5 predicates derived from the 3 available attributes; the age attribute gives the predicates [age = *child*], [age = *adult*], and [age = *old*], while the remaining Boolean attributes give the predicates [*mutation*] and [*asthma*]. Each of these can be either included pristine, included negated, or not included at all, resulting in $5^3 = 243$ combinations in total, each of which corresponds to a subgroup.

	Subset Description	f_{exc} $\bar{y}_Q - \bar{y}_E$	f_{gen} $\lvert Q \rvert$	Angle θ	f_{gwi} $\gamma = 2/3$	Pareto Optimal
1.	[age $\neq$ *child*]	2.20	18	70.4°	8.94	✓
2.	[asthma]	3.26	10	46.5°	6.88	✓
3.	[age $=$ *old*]	2.93	9	46.5°	6.19	✗
4.	[age $\neq$ *child*] $\wedge$ [asthma]	4.26	7	29.4°	5.93	✓
5.	[age $\neq$ *child*] $\wedge \neg$[mutation]	1.04	14	77.7°	5.89	✗

Table 2.2 [Top-k Subgroups]: The top-5 subgroups for the GWI function with parameter $\gamma = 2/3$. The top-k subgroups are often not Pareto optimal, as is here the case already for 2 out of the top-5 subgroups. They also exhibit redundancy by being closely positioned in the objective space; this is quantified by the small range span by the angles of the corresponding points in the objective space, here only 29.4–70.4. Notably, their order offers no control of their angle property.

of $\gamma = 2/5$, we discover the optimal subgroup [*mutation*]; this subgroup—marked in Fig. 2.1 with a square—describes a sub-population with 6.06 days of recovery more than normal, and amounts to 19% of all patients.

Once we consider the size of the corresponding sub-population, however, we can easily imagine a scenario when this subgroup turns out not to be a good fit for the intended purpose. For instance, when our goal is to introduce rules for the protection of the associated vulnerable sub-population, covering only 19% of the patients might be too specific and leave out important portions of the population. On the other hand, when our goal is to allocate limited resources, such as intensive care beds, the described sub-population might be forbiddingly populous, and therefore a less general one would be preferable, instead. In such cases a different relation between the two terms is required; to achieve exactly this, one convenient way is to tweak the γ parameter of the GWI function, which we can already see in a first overview of the subgroups that are optimal for different values of this parameter, as listed in Table 2.1. Indeed, if we emphasise on generality we discover the subgroup [age $\neq$ *child*] that describes 67% of the patients, while emphasising on exceptionality yields the subgroup [mutation] $\wedge$ [asthma] that refers to a more specific sub-population of only 15% of the patients. From the statistical perspective, the assumptions dictating a given fixed relation between these two terms might be erroneous or inaccurate, thereby misleadingly indicating that the sole discovered subgroup is truly the most significant.

Under this light, it becomes useful that the algorithm provides a collection of subgroups which are each optimal for varying relations of exceptionality

and generality, instead of just the optimal of a fixed relation between the two. A common yet problematic attempt to obtain a collection of subgroups is to compute the top-k subgroups for a given objective function. To see this in practice, we mark in the objective space (see Fig. 2.1) the top-5 subgroups of our example for $f_{\text{gwi}}(\cdot; 2/3)$ with circles and also list them in Table 2.2. A first issue of this approach is that one of the top-k subgroups might be worse than another subgroup in terms of both generality and exceptionality at the same time. This is the case already for 2 out of the 5 subgroups in the top-5 listing, and more specifically for the 3$^{\text{rd}}$ and 5$^{\text{th}}$ one, which are each worse in both terms than the 7$^{\text{th}}$ and 8$^{\text{th}}$ of listing in Table 2.1, respectively. In fact, the latter listing contains all Pareto optimal[11] subgroups, which are exactly those that are not dominated by another subgroup on both terms. We hence refer to this important set as the Pareto frontier of subgroups with respect to the two terms of exceptionality and generality.

By comparing the two listings of Tables 2.1 and 2.2 we additionally observe yet another known issue with the top-k approach: the top-k subgroups can be highly redundant [vLK11], which further translates into a very similar relation between the two terms of interest. As an intuitive measure of this relation in the objective space, we introduce the angle[12] θ of the point in polar coordinates. This additionally makes evident not only that finding the top-k subgroups provides little control over this property, which can be quantifiably demonstrated by the limited range of the angle θ of the resulting subgroups. Here, this angle lies in $\theta \in [29.4 - 77.7]$ degrees for the top-5 subgroups, whereas the Pareto frontier spans the full range of 90 degrees.

In other words, the Pareto frontier of subgroups contains optima with a variety between the two importance metrics, and therefore conveys broader information for the data at hand, while guaranteeing that there is no other Pareto optimal subgroup between any two consecutive entries. This, in turn, provides further insight on how local modifications to the predicates of each description affect the relative importance of the two term; for instance, inspecting the 3$^{\text{rd}}$ Pareto optimal subgroup [`mutation`] $\wedge$ [`asthma`] we see that the best next step to increase its exceptionality is to restrict the patient age to non-children, while to get the next most general optimal subgroup we may replace the [`asthma`] restriction with an age-related one.

By computing the Pareto frontier of subgroups we essentially approach subgroup discovery as a multi-objective optimisation (MOO) problem with

[11] In this paragraph we are content with intuitive definitions of these important terms, which we complete later with their formal equivalents.

[12] For reasons that will become apparent later, we define this angle in the *logarithmically* scaled objective space: $\theta(Q) \coloneqq \tan^{-1}\left(\ln f_{\text{gen}}(Q) / \ln f_{\text{exc}}(Q)\right)$.

the exceptionality and generality as its two objectives. However, the Pareto optimal subgroups might be inconveniently many, and a more concise subset can be useful, instead. In fact, we will later show that we can compute such a concise subset by optimising the GWI objective function over the entire range of its trade-off parameter γ, which results in this useful subset of the Pareto frontier. Hence, a key part of our second contribution is our proposal to use our algorithm of Section 2.4 to optimise the GWI objective function over $\gamma \in [0, 1]$. We additionally show that our resulting method is a MOO for subgroup discovery that selects a concise subset of these points that can be precisely described using core concepts of convex analysis. In the sequel we will first provide a broad overview of the existing MOO methods for subgroup discovery, after which we provide a formal description of the concise collection of subgroups obtained by our proposed method.

2.5.1 Related Work

Astonishingly, our treatment of subgroup discovery as a MOO problem adds to a minimally studied perspective for subgroup discovery. The existing works that adopt the MOO perspective are highly related to the relevant literature on pattern mining, and we thus start our exposition from this field, for which we distinguish primarily two classes of methods. The first consists of evolutionary methods adapted to mining rules [PMD+11; SR11], with a similar work on fuzzy subgroup description [dGH07]. These are heuristic methods that are random in nature and as such provide no guarantees of optimality—they are therefore only suitable as a first exploration of the Pareto frontier. The second class consists of *skyline* pattern mining [SCR04; UBC+17], which is a synonym coined for the Pareto frontier when applied on itemset mining. Within this setting, the state-of-the-art method for binary data [UBC+17] is able to mine a subset of the Pareto frontier specified by the user as a hyper-rectangle within the multi-objective space. This method uses a constraint satisfaction approach that prunes patterns that lie outside the specified hyper-rectangle by converting the provided constraints to known ones. For numerical data, the most relevant work is that of van Leeuwen and Ukkonen [vLU13], which provide a generic algorithm that uses arbitrary objectives; to find their Pareto frontier it traverses all possible subgroups while pruning branches of the search tree when an entropy-based measure of distance indicates that this branch contains only similar subgroups to one already considered. Finally, one approach of exceptional model mining (EMM) that is worth mentioning uses a single objective function which measures the distance of two models based on the Pareto frontier of some specified MOO problem [MCB21]. This method, however, only uses the MOO

formulation to evaluate the closeness of the exceptional model within the EMM framework, but in the end still produces the single most exceptional subgroup.

Although our proposed method can clearly be classified as a MOO approach, it differs from all the above in that it does not compute the entire Pareto frontier, but only a concise relevant subset of it. We next describe this subset and present some practical aspects for its computation. We begin our analysis with the introduction of some useful concepts from convex analysis[13] theory, and especially the study of convex polytopes.

2.5.2 Useful Concepts from Convex Analysis

A set that we will use extensively in our analysis is the **cone**, which is any set that contains every positive multiple of any of its elements. Specifically, we focus on certain convex cones[14] of particular interest. Some useful simple convex cones are the positive and negative orthants (see Fig. 2.2a [left])

$$\mathbb{R}^n_+ \coloneqq \{\mathbf{x} \subset \mathbb{R}^n \mid x_i \geq 0,\, 1 \leq i \geq n\} \tag{2.28}$$
$$\mathbb{R}^n_- \coloneqq \{\mathbf{x} \subset \mathbb{R}^n \mid x_i \leq 0,\, 1 \leq i \geq n\}\,; \tag{2.29}$$

for comparison, we also describe their open variants (see Fig. 2.2a [right])

$$\mathbb{R}^n_{++} \coloneqq \{\mathbf{x} \subset \mathbb{R}^n \mid x_i > 0,\, 1 \leq i \geq n\} \tag{2.30}$$
$$\mathbb{R}^n_{--} \coloneqq \{\mathbf{x} \subset \mathbb{R}^n \mid x_i < 0,\, 1 \leq i \geq n\}\,. \tag{2.31}$$

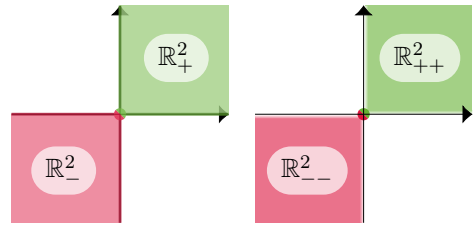
Given an arbitrary set $X \subset \mathbb{R}^n$, its **normal cone** $N_X(\mathbf{x})$ at a point $\mathbf{x} \in n$ is the set of every normal direction $\mathbf{w}$ that defines a hyper-plane through $\mathbf{x}$ that contains X in the negative half-space (see Fig. 2.2b)

$$N_X(\mathbf{x}_0) \coloneqq \{\mathbf{w} \in \mathbb{R}^n \mid \langle \mathbf{w}, \mathbf{x} - \mathbf{x}_0 \rangle \leq 0 \text{ for all } \mathbf{x} \in X\}\,, \tag{2.32}$$

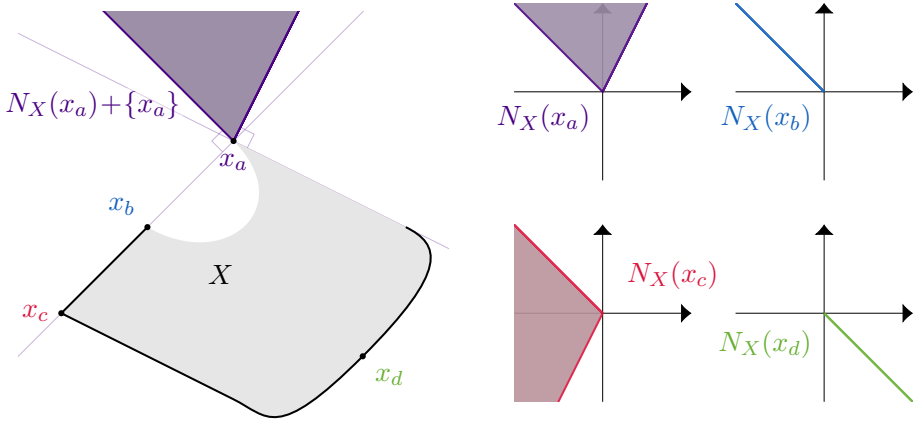
where by convention $N_X(\mathbf{x}_0) = \emptyset$ when $\mathbf{x}_0 \notin X$. The **recession cone** R_X of a set X is the cone of all **recession directions** of X (see Fig. 2.2c), that is,

[13] In this exposure we will treat real vector spaces $\mathbb{R}^n$, which are self-dual. However, we highlight the possible extensions of the following definitions and theorems to more general topological spaces, by differentiate between the sets that lie in the original and its dual space through the use of the symbols X and W, respectively. We also retain the general notation for an inner product, which allows the extension to any finite-dimensional Hilbert space by simply plugging in the associated inner product.

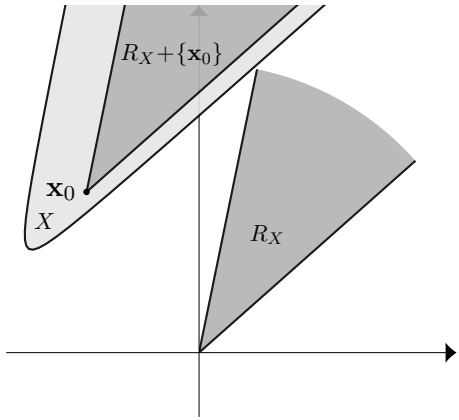
[14] Note that not all cones are convex; as a counter example consider the union of the horizontal and vertical planes in $\mathbb{R}^2$, $\{(\alpha, 0) \mid \alpha \in \mathbb{R}\} \cup \{(0, \alpha) \mid \alpha \in \mathbb{R}\}$. This set contains all positive multiples of its elements, but is clearly non-convex.



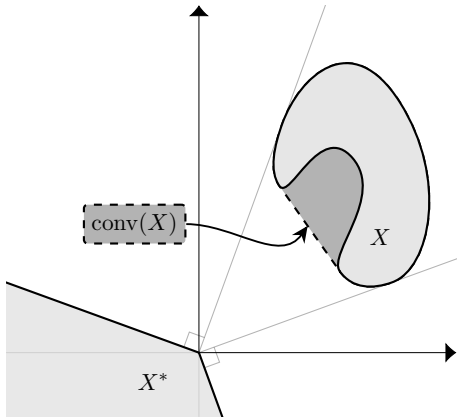
(a) The positive and negative orthants $\mathbb{R}^n_+$ and $\mathbb{R}^n_-$ [left] are simple examples of convex cones; also shown their open variants $\mathbb{R}^n_{++}$ and $\mathbb{R}^n_{--}$ [right]. Here depicted for $\mathbb{R}^2$. Note that all cones contain the origin.



(b) A non-convex set [left] and its normal cones [right] at its four points x_a, x_b, x_c, and x_d. We also depict the normal cone of point x_a shifted to have its origin at x_a, which demonstrates the orthogonality relation between the supporting hyperplanes [thin lines] of the set X and the extremal rays of its normal cone.



(c) A convex set X and its recession cone R_X. The recession cone translated at any point point $\mathbf{x}_0 \in X$ remains in the set, i.e., $R_X + \{\mathbf{x}_0\} \subseteq X$.



(d) A non-convex set X and its polar cone X^*. Also shown is the convex hull conv(X) of X, which includes in X all convex combinations of its elements (dashed area).

Figure 2.2: Visualisation of important convex cones and their key properties.

all vectors whose every positive multiple can be added to any point in the set and still result in a point of the set,

$$R_X \coloneqq \{\mathbf{x} \in \mathbb{R}^n \mid \alpha\mathbf{x} + \mathbf{x}_0 \in X, \text{ for all } \alpha \geq 0, \mathbf{x}_0 \in X\}. \tag{2.33}$$

Next, we prove a useful property of the recession cone.

Lemma 2.2. *Let set $X \in \mathbb{R}^n$ be equal to the Minkowski sum $X = S + C$ of two sets S and C, where C is a convex cone. Then the recession cone of X contains C, that is, $C \subseteq R_X$.*

Proof. Fix arbitrary points $\mathbf{x}_0 \in X$ and $\mathbf{c} \in C$. To show that $\mathbf{c} \in R_X$ we will show that the point $\mathbf{x}' \coloneqq \mathbf{x}_0 + \alpha\mathbf{c}$ also belongs to X for arbitrary $\alpha \geq 0$.

Since X is the Minkowski sum of S and C, there exist $\mathbf{x}_S \in S$ and $\mathbf{x}_C \in C$ such that $\mathbf{x}_0 = \mathbf{x}_S + \mathbf{x}_C$, so that we can write $\mathbf{x}' \coloneqq \mathbf{x}_0 + \alpha\mathbf{c} = \mathbf{x}_S + (\mathbf{x}_C + \alpha\mathbf{c})$. It now suffices to show that $\mathbf{x}_C + \alpha\mathbf{c} \in C$, as this implies that $\mathbf{x}' \in X$ by the definition of Minkowski addition for $X = S + C$. We show that as follows.

Since C is a cone it contains all positive multiples of its elements, and therefore also $2 \cdot \mathbf{x}_C$ and $2\alpha \cdot \mathbf{c}$. However, C is also convex, and therefore it contains the convex combination of its previous two points, $1/2 \cdot 2 \cdot \mathbf{x}_C + 1/2 \cdot 2\alpha \cdot \mathbf{c} = \mathbf{x}_C + \alpha\mathbf{c}$, which concludes the proof. □

Finally, the **polar cone** X^* of a set X is the set of every negative vector that corresponds to the normal direction of a supporting hyperplane for the set X (see Fig. 2.2d)

$$X^* \coloneqq \{\mathbf{w} \in \mathbb{R}^n \mid \langle \mathbf{w}, \mathbf{x}_0 \rangle \leq 0, \text{ for all } \mathbf{x}_0 \in X\}. \tag{2.34}$$

From the polar theorem we know that the polar cone is convex and closed, and it is similar to show the same for the normal cone, regardless of whether the set X has any of these properties.

Further, we denote as $\text{conv}(X)$ the **convex hull** of any set X, which contains all convex combinations of any point of X (see Fig. 2.2d). An important result for the convex hull is due to Constantin Caratheodory.

Theorem 2.3 (Caratheodory)**.** *Every point in the convex hull of a set X can be generated as the convex combination of at most $n + 1$ points from X,*

$$\text{conv}(X) = \left\{ \sum_{i=0}^{n} \alpha_i \mathbf{x}_i \;\middle|\; \mathbf{x}_0, \ldots, \mathbf{x}_n \in X \right\} \qquad \textit{for } X \subseteq \mathbb{R}^n. \tag{2.35}$$

We also use ∂X to express the boundary of X and adopt the Minkowski addition between the sets $X + X' \coloneqq \{\mathbf{x} + \mathbf{x}' \mid \mathbf{x} \in X \wedge \mathbf{x}' \in X'\}$. A simple

demonstration of Minkowski addition is the set $N_X(x_a)$ of Fig. 2.2b [left], which corresponds to translating the normal cone $N_X(x_a)$ so that its origin matches the point x_a.

2.5.3 Multi-objective Optimisation Preliminaries

We now consider the optimisation domain S of a maximisation problem with multiple objective functions that we represent as the vector-valued function $\mathbf{f} : S \to \mathbb{R}^n$. Let $X = \mathbf{f}[S]$ be the image of S through $\mathbf{f}$. In other words, $X \subseteq \mathbb{R}^n$ contains one point $\mathbf{x} \in X$ for each solution $s \in S$ in the optimisation domain, whose coordinates are equal to the value of each objective function $x_i = f_i(s)$.

Let $\mathbf{x}, \mathbf{x}' \in X$ be two points of X. We say that $\mathbf{x}$ strongly **dominates** $\mathbf{x}'$ if there is at least one dimension $1 \leq i \leq n$ for which $x_i > x'_i$ and for all other dimensions $j \neq i$ it is $x_j \geq x'_j$. Using a geometric perspective, the same requirement can be expressed as

$$\text{“}\mathbf{x} \text{ strongly dominates } \mathbf{x}'\text{”} \iff \mathbf{x}' \in \left(\{\mathbf{x}\} + \mathbb{R}^n_{--}\right) \setminus \{\mathbf{x}\} \iff \mathbf{x} \in \left(\{\mathbf{x}'\} + \mathbb{R}^n_{+}\right) \setminus \{\mathbf{x}'\}\,. \quad (2.36)$$

Albeit of lesser importance in our work, for the sake of completion but also to serve as comparison we also mention the concept of weak domination,

$$\text{“}\mathbf{x} \text{ weakly dominates } \mathbf{x}'\text{”} \iff \mathbf{x}' \in \{\mathbf{x}\} + \mathbb{R}^n_{-} \iff \mathbf{x} \in \{\mathbf{x}'\} + \mathbb{R}^n_{++}\,. \quad (2.37)$$

Using these notions, we now formally define the **Pareto optimal** points[15] of the scores in X as those of its points that are not strongly dominated by some other point in the set

$$F_X \coloneqq \left\{ \mathbf{x} \in X \;\middle|\; \{\mathbf{x}\} + \mathbb{R}^n_{++} \cap X = \emptyset \right\}. \quad (2.38)$$

The **Pareto frontier** of the multi-objective optimisation problem of maximising $\mathbf{f}$ over S is the set of Pareto optimal points of $X = \mathbf{f}[S]$, that is of the solution scores X of the solutions S with respect to the vector-values objective function $\mathbf{f}$. Such a Pareto frontier can be seen in Fig. 2.3b.

[15] For completion we also mention that one can also define the strong Pareto frontier of X to be the points in X that are not weakly dominated by any other points of this set, $\left\{ \mathbf{x} \in X \;\middle|\; \{\mathbf{x}\} + \mathbb{R}^n_{+} \cap X = \{\mathbf{x}\} \right\}$.

Finally, of special interest is the **support function** of an arbitrary set,

$$\mathrm{supp}_X(\mathbf{w}_0) \coloneqq \sup_{\mathbf{x}\in X} \langle \mathbf{w}_0, \mathbf{x}\rangle \,, \tag{2.39}$$

which for $\|\mathbf{w}_0\| = 1$ gives the distance from the origin of the supporting hyperplane with normal $\mathbf{w}_0$ that contains X in its negative[16] half-space. In this way it defines the **supporting hyper-plane** $H_{\mathbf{w}_0}$ and corresponding (negative) **half-space** $L_{\mathbf{w}_0}$

$$H_{\mathbf{w}_0} \coloneqq \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{w}_0, \mathbf{x}\rangle = \mathrm{supp}_X(\mathbf{w}_0)\} \tag{2.40}$$
$$L_{\mathbf{w}_0} \coloneqq \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{w}_0, \mathbf{x}\rangle \leq \mathrm{supp}_X(\mathbf{w}_0)\} \tag{2.41}$$

2.5.4 Computing a Concise Subset of the Pareto Frontier

We can now provide a formal characterisation of the particular subset of the Pareto frontier that our method computes. We start with an general important theorem.

Lemma 2.4 (Pareto frontier duality)**.** *Let $X \subseteq \mathbb{R}^n$ be a closed set and define $g(\mathbf{w}_0) \coloneqq \arg\max_{\mathbf{x}\in X}\langle \mathbf{w}_0, \mathbf{x}\rangle$ to be the maximiser of its support function and $C \coloneqq \mathrm{conv}(F_X + \mathbb{R}^n_-)$ the convex hull of its extension with the negative orthant. Also, define $B \coloneqq \partial C \cap F_X$ to be the Pareto optimal points of X that lie on the boundary of C, and let $W \coloneqq R^*_C$ be the polar cone of the recession cone of C.*

1. *Then the following dual relationships hold:*

$$\mathbf{x} \in g(\mathbf{w}) \quad \Longrightarrow \quad \mathbf{w} \in N_C(\mathbf{x}) \wedge \mathbf{x} \in B \qquad \text{for all } \mathbf{w} \in W \tag{2.42}$$
$$\mathbf{w} \in N_X(\mathbf{x}) \quad \Longrightarrow \quad \mathbf{x} \in g(\mathbf{w}) \wedge \mathbf{w} \in W \qquad \text{for all } \mathbf{x} \in B \tag{2.43}$$

2. *The maximisers of the support function coincide with the set of all Pareto optimal points along directions W that also lie on the boundary of the convex hull of X*

$$\bigcup_{\mathbf{w}\in W} g(\mathbf{w}) = B \,. \tag{2.44}$$

Proof. We start with part 1 and show Eq. (2.42).

[16] We note that in its usual definition, the supporting hyper-plane contains the supported set in the positive half-space. We here allow this deviation for the sake of notational simplicity.

We let $\mathbf{w} \in W$ and $\mathbf{x} \in g(\mathbf{w})$, and we first show that $\mathbf{w} \in N_X(\mathbf{x})$. By the definition of the normal cone in Eq. (2.32), it suffices to show that for arbitrary $\bar{\mathbf{x}} \in C$ it is $\langle \mathbf{w}, \bar{\mathbf{x}} - \mathbf{x} \rangle < 0$.

Since C is the convex hull of $F_X + \mathbb{R}^n_-$, by Theorem 2.3 there exist[17] $n+1$ points $\mathbf{x}_0, \ldots, \mathbf{x}_n \in F_X + \mathbb{R}^n_-$ such that

$$\bar{\mathbf{x}} = \sum_{i=0}^{n} \alpha_i \mathbf{x}_i \text{ with } \sum_{i=0}^{n} \alpha_i = 1 \qquad \text{where } \alpha_i \geq 0,\, \mathbf{x}_i \in F_X + \mathbb{R}^n_- \,. \tag{2.45}$$

Within this expression, each of the constituent points $\mathbf{x}_i$ belongs to the Minkowski sum of F_X and $\mathbb{R}^n_-$, and each can therefore be written as

$$\mathbf{x}_i = \mathbf{y}_i + \mathbf{z}_i, \qquad \text{with } \mathbf{y}_i \in F_X,\, \mathbf{z}_i \in \mathbb{R}^n_-, \qquad \text{for all } 0 \leq i \leq n \,. \tag{2.46}$$

However, by the original assumptions it is $\bar{\mathbf{x}} \in g(\mathbf{w}) \coloneqq \arg\max_{\mathbf{x}' \in X} \langle \mathbf{w}, \mathbf{x}' \rangle$, and since $\mathbf{y}_i \in F_X \subseteq X$ we have

$$\langle \mathbf{w}, \bar{\mathbf{x}} \rangle \geq \langle \mathbf{w}, \mathbf{y}_i \rangle \iff \langle \mathbf{w}, \mathbf{y}_i - \bar{\mathbf{x}} \rangle \leq 0, \qquad \text{for all } 0 \leq i \leq n \,. \tag{2.47}$$

By Lemma 2.2 applied on the set C, the recession cone R_C contains the negative orthant $\mathbf{z}_i \in \mathbb{R}^n_- \subseteq R_C$. However, since $\mathbf{w} \in W \coloneqq R_C^*$, by the definition of the polar cone of Eq. (2.34) the vectors $\mathbf{w}$ and $\mathbf{z}_i$ are related by the equation

$$\langle \mathbf{w}, \mathbf{z}_i \rangle \leq 0, \qquad \text{for all } 0 \leq i \leq n \,. \tag{2.48}$$

We can now add both sides of Eqs. (2.47) and (2.48) to get

$$\langle \mathbf{w}, \mathbf{y}_i - \mathbf{x} \rangle + \langle \mathbf{w}, \mathbf{z}_i \rangle \leq 0 \iff \langle \mathbf{w}, \bar{\mathbf{x}}_i - \mathbf{x} \rangle, \text{for all } 0 \leq i \leq n \,; \tag{2.49}$$

we now multiply each of the inequalities in Eq. (2.49) by $\alpha_i > 0$ and then sum them all to get

$$\sum_{i=0}^{n} \alpha_i \langle \mathbf{w}, \mathbf{x}_i - \mathbf{x} \rangle \leq 0 \iff \langle \mathbf{w}, \sum_{i=0}^{n} \alpha_i \mathbf{x}_i - \mathbf{x} \cancel{\sum_{i=0}^{n} \alpha_i} \rangle \leq 0 \iff \langle \mathbf{w}, \bar{\mathbf{x}} - \mathbf{x} \rangle \leq 0 \quad \text{for all } 0 \leq i \leq n \,, \tag{2.50}$$

[17] Although the Caratheodory theorem allows $\mathbf{x}'$ to be described by less than $n+1$ points, we use here exactly $n+1$ for simplicity and without loss of generality; indeed, if less than $n+1$ points are needed, we can simply repeat a needed point multiple times and set its coefficient to 0, until we have a convex combination of exactly $n+1$ points.

which is what we needed to show. □

We can now apply the intuition of Lemma 2.4 to develop a MOO optimisation framework for subgroup discovery. To this end we introduce the vector-valued function $\mathbf{f}_{\text{lin}} : 2^E \to \mathbb{R}^2$, each dimension of which corresponds to each of the terms in the objective function f_{gwi} of Eq. (2.10). Namely, $\mathbf{f}_{\text{lin}}$ maps every subset $Q \subseteq E$ to the pair of values corresponding to its size and its mean deviation, $\mathbf{f}_{\text{lin}} : Q \mapsto (|Q|, \bar{y}_Q - \bar{y}_E)$. Although this vector-valued objective is still useful, we will further need its logarithmically scaled variant, $\mathbf{f}_{\text{gwi}} : 2^E \to \mathbb{R}^2$, for which $\mathbf{f}_{\text{gwi}} : Q \mapsto \big(\log(|Q|), \log(\bar{y}_Q - \bar{y}_E)\big)$.

The importance of introducing the $\mathbf{f}_{\text{gwi}}$ is revealed by the following observation. Consider that we use the theory of Section 2.4 to compute the maxima of f_{gwi} over all extensions of the subgroups allowed by our chosen language $\text{ext}(\mathcal{L}; E)$, and then apply $\mathbf{f}_{\text{gwi}}$ over the set of maximisers. We can show that this yields an optimisation problem that is equivalent to maximising the support function of the image $\mathbf{f}_{\text{gwi}}[\text{ext}(\mathcal{L}; E)]$ of $\text{ext}(\mathcal{L}; E)$ through $\mathbf{f}_{\text{gwi}}$:

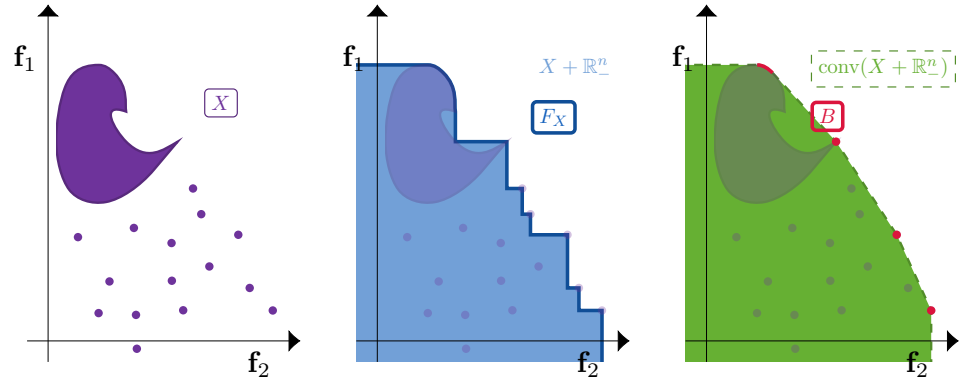
$$\mathbf{f}_{\text{gwi}}\left[\operatorname*{arg\,max}_{Q \in \text{ext}(\mathcal{L};E)} |Q|^{\gamma}(\bar{y}_Q - \bar{y}_E)^{1-\gamma}\right] = \mathbf{f}_{\text{gwi}}\left[\operatorname*{arg\,max}_{Q \in \text{ext}(\mathcal{L};E)} \gamma \log(|Q|) + (1-\gamma)\log(\bar{y}_Q - \bar{y}_E)\right] = \operatorname*{arg\,max}_{\mathbf{x} \in \mathbf{f}_{\text{gwi}}[\text{ext}(\mathcal{L};E)]} \langle \mathbf{w}_\gamma, \mathbf{x} \rangle\,, \quad (2.51)$$

where $\mathbf{w}_\gamma = \frac{1}{\sqrt{\gamma^2 + (1-\gamma)^2}}(\gamma, 1-\gamma)^\top$ is the unit-length normal direction that corresponds to a particular choice of the γ parameter. We can now apply Lemma 2.4 for the special case of the set $\mathbf{f}_{\text{gwi}}[\text{ext}(\mathcal{L}; E)]$ to show the structure of the optima that result from the procedure of sweeping $\gamma \in [0, 1]$.

Corollary 2.5. *Denote $S \coloneqq \text{ext}(\mathcal{L}; E)$ the set of allowed subgroups. Then the set of optimal subgroups discovered by maximising f_{gwi} over S for each $\gamma \in [0, 1]$ coincides with the set of all strongly Pareto optimal subgroups whose multi-objective value vector lies on the boundary of the convex hull of the image $X_{\text{gwi}} \coloneqq \mathbf{f}_{\text{gwi}}[S]$:*

$$\bigcup_{\gamma \in [0,1]} \operatorname*{arg\,max}_{Q \in S} f_{gwi}(Q; \gamma) = \partial\, \text{conv}(X_{\text{gwi}} + \mathbb{R}^n_-) \cap F_{X_{\text{gwi}}}\,. \quad (2.52)$$

Of course, this result does not limit itself only to f_{gwi}, but to any geometrically weighted function and of any number of terms, after we simply replace X_{gwi} with the image of the optimisation domain through the new



(a) The image $X = \mathbf{f}[S]$ of an example of solution domain S through a 2-dimensional multi-objective function $\mathbf{f}$.

(b) The Pareto frontier of $\mathbf{f}$ [thick line] and the expansion of X with the directions in the negative orthant $X + \mathbb{R}^n_-$ [filled area].

(c) The convex hull of $X + \mathbb{R}^n_-$ and the result set B of Eq. (2.44).

Figure 2.3: Depiction of some useful sets that are described in Lemma 2.4.

function. In fact, all of the objective functions that we will study in this work fall under this description, and we will extensively use Corollary 2.5 to characterise the optima of our methods. What is more, in each of the following chapters we will present the results of the corresponding methods over a broad range of a parameter that will serve as an equivalent to the γ parameter that we presented here. As a result, all subgroup discovery methods in this thesis are instances of this MOO framework for subgroup discovery.

An additional result that will allow us to simplify the results of the following sections is expressed as the following corollary of Lemma 2.4.

Corollary 2.6. *Let* $\mathbf{x} \in g(\mathbf{w})$ *and* $\mathbf{x}' \in g(\mathbf{w}')$ *for* $\mathbf{w}, \mathbf{w}' \in W$ *with* $\mathbf{w} \neq \mathbf{w}'$. *Then there is no extremal point of* $\operatorname{conv}(S + \mathbb{R}^n_-)$ *in the set* $\bigcup_{\alpha \in (0,1)} g(\alpha \mathbf{w} + (1-\alpha)\mathbf{w}')$.

Using Corollary 2.6 we know that if two γ values yield the same optimiser, then no other optimiser will be found for any γ between the former two, up to equivalent subgroups. Practically, we can therefore merge all intervals of γ at the endpoints of which we find the same subgroup.

When we apply our method to our toy example, we see that out of the full Pareto frontier we only select the a smaller subset of subgroups, as shown in Fig. 2.4. This set comprises those conspicuous points that are Pareto optimal, but also correspond to extremal points of the convex hull of the

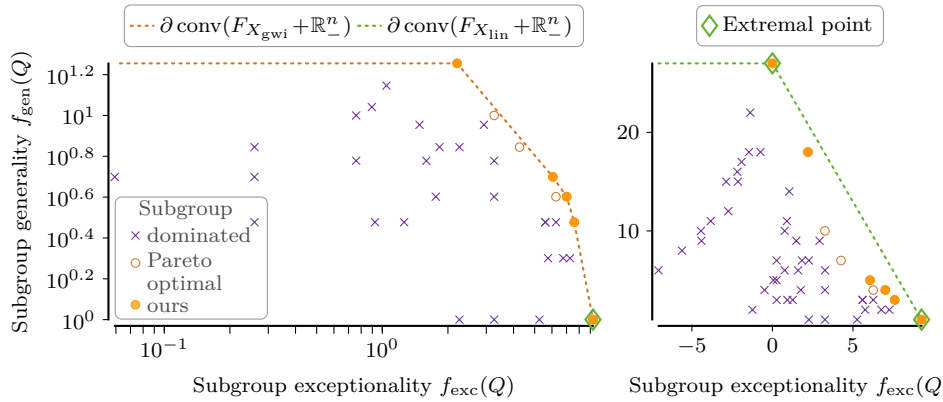


Figure 2.4 [Our Concise Subset of Pareto Optima]: The Pareto frontier (empty circles) of the objective space from our toy dataset (see Fig. 2.1), shown both in dual logarithmically scaled axes [left] and in linear, un-scaled ones [right]. We also highlight the subset of the Pareto frontier selected by our method (filled circles). This set includes only extrema of the convex hull of $F_{X_{\mathrm{gwi}}}+\mathbb{R}^n_-$ and leaves out subgroups which are Pareto optima, but offer nevertheless minimal improvement on other included points; this results in a more concise set of subgroups that comprises the subset of the conspicuous Pareto optima. Compared to our approach, when no logarithmic scaling is used, the set of extremal points of the corresponding convex hull contains only the two trivial subgroups that correspond to the entire population and the single most extremal point, respectively.

set $F_X + \mathbb{R}^n_-$. Importantly, the logarithmic scaling of the two objectives, as implied by our the choice of the geometrically weighted f_{gwi}, has the effect that the subset of extremal points is populous enough to capture important subgroups. In contrast, if we were to use simpler objective functions that would result in a weighted arithmetic mean, which is a common approach for general MOO [CH08], we would only be able to find a much smaller subset of these points. Similar to the argumentation of Corollary 2.5, this subset would correspond to the extrema of the set $\mathrm{conv}(X_{\mathrm{lin}}+\mathbb{R}^n_-)$, where we denote $X_{\mathrm{lin}} \coloneqq \mathbf{f}_{\mathrm{lin}}[S]$ the image of allowed subgroups through the un-scaled vector-valued objective function $\mathbf{f}_{\mathrm{lin}}$. We depict this subset in Fig. 2.4 [right], from which we can see that, in our example, not only corresponds to only two subgroups, but in fact also to the most trivial ones: the entire dataset and one containing just the single, most extreme point, and a potential outlier.

Overall, our proposed approach forms a more concise subset that retains the salient subgroups out of the full Pareto frontier, the latter of which can often be forbiddingly populous. Due to these benefits, in the following we

will always adopt the multi-objective approach for treating our results, and will always report subgroups for a varying relation between the two scores of interest.

This concludes our extensive overview and formal definition of subgroup discovery and the principles motivating the choice of objective functions, including our proposed MOO framework. In the following sections we will use the novel algorithm that we presented in Section 2.4.2 to solve several of the limitations with typical subgroup discovery that we presented in Chapter 1.

3 Representative Subgroups

> The very fact that racism degrades both the perpetrator and the victim commands that, if we are true to our commitment to protect human dignity, we fight on until victory is achieved.
>
> *(Nelson Mandela, United Nations General Assembly address)*

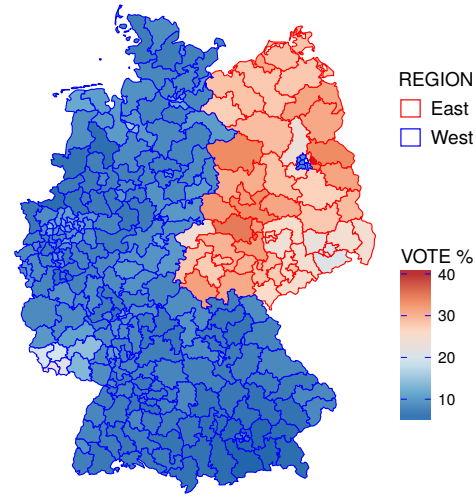
While typical subgroup discovery can find the most exceptional entity subgroup, this goal is not always sufficient on its own. In several practical scenaria, the data might be affected by trends and biases to an extent that the most exceptional subgroup can overshadow interesting detail in the population, or under-represent important sub-populations in the data. Then, the discovered subgroup can range from unimportant to completely uninformative, or even reflect an unacceptably unfair view of the data. We first encountered these two scenaria in Section 1.3.1 while studying the sensitivity of a novel dye on a toy dataset of bacterial species (see Fig. 1.2). The issue of one prominent sub-population overshadowing important novel discovery commonly arises in real-world settings from scientific discovery and theory development, when we seek to identify local factors that influence some variable while desiring to control the influence of other potential explanations or contributing factors. In other cases, we need to control the often unavoidable bias in the data and prevent it from being reflected in the found sub-populations, so that we ensure a fair *representation* of any sensitive attribute, such as gender or race.

A realistic setting where trends in the data might overshadow important subsets comes from an application in materials science, in which a common task is to discover structural properties that characterise the difference in the energy between the electrons of the highest occupied molecular orbital (HOMO) and the lowest unoccupied one (LUMO). This so-called HOMO-LUMO energy gap is a very decisive indicator of many physical and chemical properties for many important families of materials. One such family with broad biomedical [KAS+18] and other applications are gold nanoclusters, i.e., materials whose molecules consist solely of gold atoms and form a

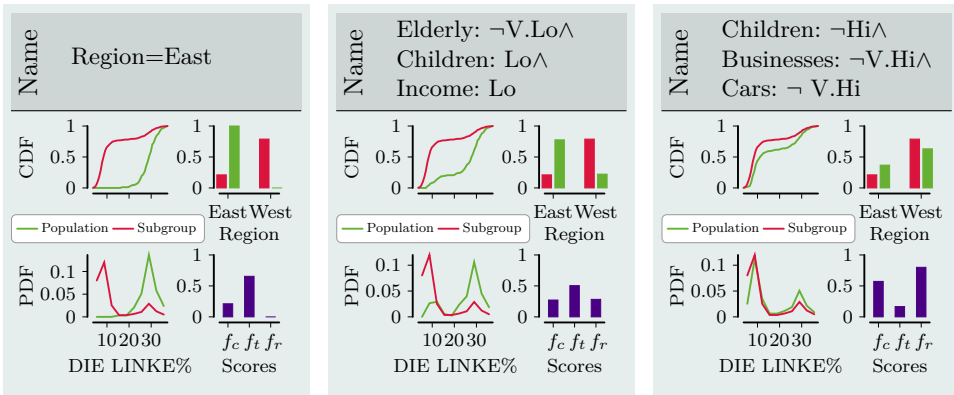
characteristic arrangement in space with a given number of atoms. Since the properties of these materials are greatly influenced by the parity of the atom number, whenever we apply typical subgroup discovery on any of these properties as a target concept, it simply discovers the subgroup of atoms with either even or odd parity [GBV+17], which is truly an exceptional subgroup. However, since the domain expert is already aware of the effects of atom number parity, this subgroup is of little use. One good way to go beyond this already known trend in the data is to steer the algorithm toward subgroups which describe the factors that equally affect nanoclusters of both even and odd number of atoms—that is, subgroups that are *representative* for both parity classes.

As another example, in political science we are often interested in discovering demographics with a high affinity to a certain political party. However, when studying the popularity of the left party during the 2009 German parliament elections (see Fig. 3.1), it becomes evident that there is a strong influence of the geographical location of each voting district (see Fig. 3.1a) due to several factors, including the common underlying historical coherence of each region. When typical subgroup discovery is used, it recovers the subgroup [`region` = `East`], which, despite being exceptional and accurate, offers negligible new insight. What is more, the naïve workaround of removing the region information from the data yields a similar sub-population, despite the fact that its description does not involve the region attribute. The way we can resolve this problem is to require that the descriptions we find represent both regions to the same extent. Only then do we recover the common underlying factors that affect voting for the left party in both regions, beyond the already known factor of geography: "*voters with few children and not very high number of businesses or cars*" (see Fig. 3.1d).

From a different perspective, the usefulness of representative subgroups goes beyond the rejection of known trends; in fact, it is the same mechanism that can deliver subgroups that remain balanced and therefore fair, whereas traditional subgroup discovery may inadvertently pick sub-populations with unfair composition with respect to a sensitive trait, like race or gender. In applications like policy development and other fairness-sensitive domains, ethical and legal requirements might call for the distribution of policy recipients to not under-represent individuals of a minority, or to equally apply to all classes of a sensitive trait. Consider, for example, that a committee for the selection of student applicants seeks common underlying factors in their previous alumni that led to professional success, in order to establish rules for the granting of a scholarship. Let us suppose in a not-so-hypothetical setting that the various biases in the society introduce a disproportionate



(a) The historical membership of each voting district to eastern or western Germany strongly influenced the voting outcome of the left party "DIE LINKE" during 2009 German elections. District region memberships are marked with coloured outlines.



(b) Typical **(c)** Ignoring "*region*" **(d)** Representative

Figure 3.1: Subgroups of German voting during 2009 elections with high percentage of the left-wing party "DIE LINKE", which is known to be strongly influenced by the geography (a). Each block shows the cumulative distribution (CDF) and probability density (PDF) over all voting districts for each vote percentage in both the subgroup (red) and the global population (blue) [left]. Also shown is the distribution of district locations [top right] and the score of the coverage f_c, tendency f_t, and representativeness f_r terms of the objective function, respectively [bottom right]. Traditional subgroup discovery (b) recovers the main trend: eastern districts support "DIE LINKE". Removing the `region` attribute (c) results in a similar subgroup. Only when explicitly controlling for geography (d) do we discover subgroups that stand out with regard to voting behaviour, while at the same time being representative for the whole country.

amount of successful white males in the data, when contrasted to their non-white or female counterparts [RC19]. Should a traditional subgroup discovery method be used here, the sub-populations that it would return as worthy of a scholarship would have a large overlap with the sub-population of white males. Arguably, this contradicts the purpose of scholarships, which should ideally be to level the field of opportunity based solely on the abilities of the applicants, instead of their disproportionate availability of past resources [19]. This therefore necessitates for algorithms that would enforce that both genders or all races would be equally represented in the discovered sub-populations, that should serve as the eligible for the scholarship.

This very requirement arises in numerous cases where we want to fulfil fairness guidelines with respect to some sensitive attribute. In this case, the constraint that all classes of this attribute must occur with the same frequency is called *statistical parity* [ZWS+13] and constitutes the first and foremost approach in mending the inherent bias in the data. Importantly, the effect we discussed in the application of election analysis above, in which simply ignoring or removing the sensitive attribute from the data fails to guarantee representativeness, can also appear in fairness applications. In this setting, recovering the sensitive variable by as a combination of the non-sensitive variables that remain in the dataset is known as the *red-lining effect* [CV10]. This effect is a well-studied artefact of fairness-unaware [DHP+12] algorithms, but also underlies the deliberate practice of gerrymandering, which is an unethical means of gaining an substantial unfair advantage during elections by carefully carving the voting districts [Sof16; Ste17b].

Altogether, in all the above settings, we can express the common requirement as being statistically *representative* in relation to some given sensitive attribute of the population: that would be the Gram classification in the dye example, the parity of atom number in the study of nanoclusters, the region in the case of voting analysis, or the racial and gender identification of a scholarship applicant. However, little work has been previously done on this task. From the perspective of overcoming known trends in the data, other approaches have focused on the notion of unexpectedness with respect to background information [GMM+07; MVT12], where the new discovery must be surprising when compared to a set or structures that are means to capture prior information; none of these methods, however, can guarantee that the results will be representative. From the fairness perspective, this requirement has been more widely enforced, with methods ranging from data pre-processing, model adaptation and post-processing schemes [DL19]. Out of these, the former come with their own difficulties: e.g.: non-transparent data modification or aggressive removal of entities, and post-processing tech-

niques are not as efficient from the optimisation perspective of subgroup discovery. When we now focus on the rest, there exist no methods that provide any local models, akin to subgroups, such as the one we desire, and when it comes to standard subgroup discovery it is neither clear how to i) generalise to settings that go beyond a binary prediction task, or ii) how to use in an efficient method for optimal subgroup discovery.

Therefore, in this chapter we first introduce the concept of a **control variable**: an additional variable defined for each entity in the dataset, that represents either the sensitive trait or, more generally, a classification variable whose distribution within the resulting subgroup we wish to be close to a given prototypical distribution. Based on this control variable, we then propose a representativeness term that extends the impact function of Eq. (2.9), which ensures that the sensitive trait within the subgroup matches a pre-defined, desired distribution. We then use this objective to form a method that solves the task of finding representative subgroups efficiently. In particular, we propose RAWR, an algorithm to compute the tight optimistic estimator for the representativeness-aware objective function in $O(n \log n)$ time for the popular case of a binary control and a numeric target variable, and that can be used in our IDDFS framework (see Section 2.4.2). What is more, our objective function introduces a similar weighting parameter to the GWI of Eq. (2.10), which we can therefore use to traverse the subset of the Pareto frontier defined by the two objectives of i) representativeness and ii) typical exceptionality, according to the theory we developed in Section 2.5.

On performed experiments in real world data, we show that the pruning capacity of our specialised optimistic bound for the novel task of finding representative subgroups is superior by orders of magnitude than the next best possible alternative. Alongside the minimal memory footprint arising as a direct result of the use of our branch-and-bound framework (see Section 2.4.2), we show that the pruning superiority of our specialised objective function allows similar improvements in running times, and therewith RAWR makes it possible to optimally compute the most representative subgroup in an otherwise computationally infeasible setting. Importantly, we also qualitatively study the usefulness of the discovered sub-populations, with several real-world settings that demonstrate the potential impact of this method.

3.1 Measuring Subgroup Representativeness

We start by introducing some useful notation, before we proceed to formalise our intuition into a measure of representativeness. We then incorporate this

measure to form an objective function for assessing exceptional yet representative subgroups; we then will present RAWR, an efficient implementation of the tight optimistic estimator for our objective function, applicable for balanced binary controls.

For the needs of the proposed task of **representative subgroup discovery**, we assume that an additional **control variable** $c : E \to \{1, \ldots, K\}$ is defined on each entity, similar to the target variable $y : E \to \mathcal{V}_y$ of typical subgroup discovery. This control variable represents a sensitive class whose distribution within the subgroup we desire to be close to a prototypical one. Our research goal, therefore, becomes as follows.

Goal G4. *Find the subgroup $Q \in \mathcal{L}$ with*

- *the most exceptional distribution of the target variable, in which*
- *the distribution of the control variable is close to a prototypical one.*

We also impose a total ordering over each entity subset $Q \subseteq E$, by ordering its entities decreasingly with respect to their target value, so that $\epsilon_i \leq \epsilon_j \iff y(\epsilon_i) \geq y(\epsilon_j)$. Hence, $\epsilon_{\mathrm{i}}^{(Q)}$ is the entity within Q with the i-th greatest target value, whose target value itself we denote $y_{\mathrm{i}}^{(Q)}$ and its control value $c_{\mathrm{i}}^{(Q)}$. Out of those elements of Q with class k, we denote $y_{\mathrm{i}}^{(k)}$ the one with the i-th greatest target value, and by $n_k(Q) \coloneqq |\{\epsilon \in Q : c(\epsilon) = k\}|$ their count, which we also refer to as the k-th **class count**. Similarly, we define the **class probability vector** $\boldsymbol{p}(Q)$ with elements the **class probabilities** $p_k(Q) \coloneqq {}^{n_k(Q)}/_{|Q|}$, for each class k.

For the sake of simplicity, but also since it will be of special interest in the development of an efficient algorithm, we first focus on the case of a binary control variable $\mathcal{V}_c = \{0, 1\}$, and we later discuss extensions for the general case. Additionally, we assume a continuous **target variable**, that is, we set its domain to be $\mathcal{V}_y = \mathbb{R}$, although the same analysis also holds unmodified for the discrete case.

3.1.1 The Controlled Impact Function

We now augment the standard objective function of Eq. (2.9) to also account for a broader notion of generality than coverage: the statistical generality of the subgroup w.r.t. the control variable. Specifically, we add a **representativeness factor** $f_{\mathrm{r}}(Q)$, quantifying the similarity of the control distribution between Q and E. This forms the **controlled impact function**

$$f_{\mathrm{cif}}(Q) \coloneqq f_{\mathrm{ct}}(Q)^{1-\gamma} f_{\mathrm{r}}(Q)^{\gamma} \ , \tag{3.1}$$

where the $\gamma \in [0, 1)$ parameter tunes the trade-off between representativeness and the typical properties quantified by f_{ct}.

To be representative, the subgroup must contain—to the extent possible—each class with the same probability as in the original dataset. As motivated from the fairness perspective of our task, the statistical parity requirement in our goal implies that we select subpopulations independently of the control variable. This is equivalent to requiring that a random entity $\epsilon \in E$ from the population satisfies

$$\mathbb{P}\big(c(\epsilon)|\sigma(\epsilon) = \top\big) = \mathbb{P}\big(c(\epsilon)\big) \iff d(\boldsymbol{q}, \boldsymbol{p}) = 0 \,, \tag{3.2}$$

where d is some distance measure between distributions with $\boldsymbol{q} \coloneqq \boldsymbol{p}(Q)$ and $\boldsymbol{p} \coloneqq \boldsymbol{p}(E)$. In this work, we further fix d to be the **total variation distance** $d(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2} \sum_k |q_k - p_k|$, equal to the maximal difference between probabilities of any set of control classes. This measure is at once intuitively interpretable and simple enough to allow for its efficient computation, and we therefore adopt it as our choice throughout this work.

For a better comparison of the relevant values among different datasets, we hence normalise all terms to obtain values in the interval $[0, 1]$, which thus gives the form of the representativeness factors as:

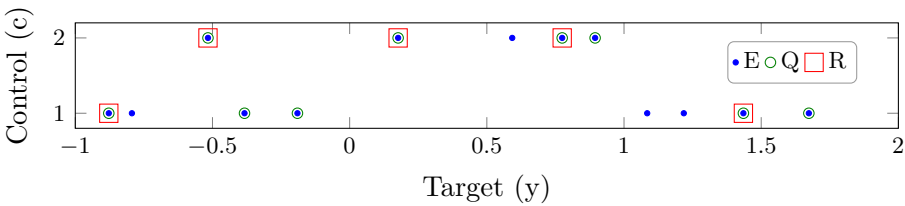
$$f_{\mathrm{r}}(Q) \coloneqq 1 - \frac{d(\boldsymbol{p}, \boldsymbol{q}) - d_{\max}}{d_{\max}}, \qquad d_{\max} \coloneqq \max_{R \subseteq E} d(\boldsymbol{p}, \boldsymbol{r}) \,. \tag{3.3}$$

An important consequence of this normalisation of the representativeness term is that any optimistic estimator for the impact function is also a valid, albeit non-tight, optimistic estimator for the controlled impact function.

Having introduced all constituents of the controlled impact function, we now proceed with the computation of its tight optimistic estimator. We first introduce a transform of the domain of the original optimisation problem from exponential to polynomial size in Section 3.1.2. We then employ this transform in Section 3.1.3 to derive an efficient algorithm that computes this tight optimistic estimator in $O(n \log n)$ time, for the special case of a population with balanced binary classes.

3.1.2 Efficient Searching in the Class Counting Space

Next, we describe a transformation which aggregates the exponentially many subsets of Q in the original optimisation problem of Eq. (2.27) into polynomially many sets of subsets. Additionally, the maximum f_{cif} value attained by any subset within each of these sets can be efficiently computed. We recall that due to their intended use within the branch-and-bound



(a) Toy population E, consisting of $n_1(E) = 8$ items of control class 1 and $n_2(E) = 5$ of class 2, a subgroup $Q \subseteq E$, with $n_1(Q) = 5$ items of class 1 and $n_2(Q) = 4$ of class 2, and a refinement $R \subset Q$.

(b) The CCS of Q, denoted $\mathcal{I}(Q)$, and the maximum ${f_r}^Q$ ray. Each refinement $R \subseteq Q$ with class counts $\boldsymbol{I} = (I_1, I_2)$, where $I_1 = n_1(R)$ and $I_2 = n_2(R)$, is contained in the equi-class refinement set $\mathcal{R}_{\boldsymbol{I}}$, which corresponds to the point $\boldsymbol{I}$ in $\mathcal{I}(Q) = \{0, \ldots, 5\} \times \{0, \ldots, 4\}$. Points closer to the *max* f_r *ray* have a class count probability (ratio) closer to that of E and thus a higher f_r score.

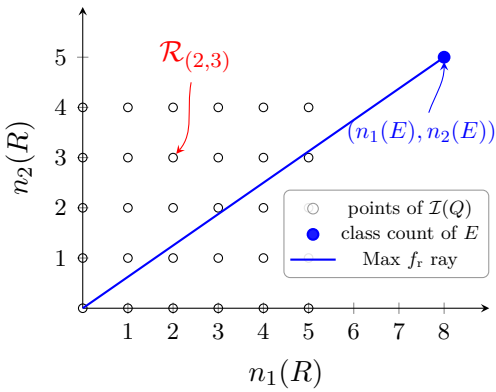


Figure 3.2: The class counting space (bottom) for a toy population with $K = 2$ control classes (top). The refinement R is contained in $\mathcal{R}_{(2,3)}$, corresponding to the annotated point.

framework, all subsets of $R \subseteq Q$ are called **refinements** of Q.

For any given subgroup Q, we consider the space of all possible class count vectors $\boldsymbol{I} \coloneqq (n_1(Q), \ldots, n_K(Q))$ that any refinement $R \subseteq Q$ might assume,

$$\mathcal{I}(Q) \coloneqq \bigtimes_{k=1}^{K} \{0, \ldots, n_k(Q)\} \ . \tag{3.4}$$

This space, which we refer to as the **class counting space (CCS)**, is a subset of the lattice $\mathbb{Z}^K$, and partitions the original space 2^Q into $|\mathcal{I}(Q)| = \prod_{k=1}^{K}(n_k(Q) + 1)$ partitions. We call these partitions the **equi-count refinement sets** $\mathcal{R}_{\boldsymbol{I}}(Q)$, each of which consist of these refinements of Q with I_k items of control class k, for each class $k = 1, \ldots, K$,

$$\mathcal{R}_{\boldsymbol{I}}(Q) \coloneqq \{R \subseteq Q : n_k(R) = I_k, \quad k = 1, \ldots, K\} \ . \tag{3.5}$$

For an example of a CCS with $K = 2$ classes see Fig. 5.1.

The computation of the tight optimistic estimator $\hat{f}_{\text{cif}}(Q) \coloneqq \max_{R \in 2^Q} f_{\text{cif}}(R)$ of Eq. (2.27) can now be expressed as

$$\hat{f}_{\text{cif}}(Q) \coloneqq \max_{\boldsymbol{I} \in \mathcal{I}(Q)} \max_{R \in \mathcal{R}_{\boldsymbol{I}}(Q)} f_{\text{cif}}(R) = \max_{\boldsymbol{I} \in \mathcal{I}(Q)} {f_{\text{cif}}}^{Q}(\boldsymbol{I}) \,, \tag{3.6}$$

where ${f_{\text{cif}}}^{Q}(\boldsymbol{I})$ refers to the maximal value attained over all refinements in the equi-count refinement set $\mathcal{R}_{\boldsymbol{I}}$

$${f_{\text{cif}}}^{Q}(\boldsymbol{I}) \coloneqq \begin{cases} \max\limits_{R \in \mathcal{R}_{\boldsymbol{I}}(Q)} f_{\text{cif}}(R) & \boldsymbol{I} \in \mathcal{I}(Q) \setminus \{\mathbf{0}\} \\ -\infty & \boldsymbol{I} = \mathbf{0} \end{cases} . \tag{3.7}$$

Similarly, the maxima of the impact function, central tendency and representativeness values over all refinements within $\mathcal{R}_{\boldsymbol{I}}$ are denoted ${f_{\text{ct}}}^{Q}(\boldsymbol{I})$, ${f_{\text{t}}}^{Q}(\boldsymbol{I})$ and ${f_{\text{r}}}^{Q}(\boldsymbol{I})$, respectively.

In the next proposition we derive a closed form for the optimiser of $f_{\text{cif}}(Q)$ within an equi-count refinement set $\mathcal{R}_{\boldsymbol{I}}$, which can then be used to compute ${f_{\text{ct}}}^{Q}(\boldsymbol{I})$ and thus ${f_{\text{cif}}}^{Q}(\boldsymbol{I})$.

Proposition 3.1. *The optimal value ${f_{ct}}^{Q}(\boldsymbol{I})$ is attained by the set*

$$R^*_{\boldsymbol{I}} \coloneqq \bigcup_{k=1}^{K} \left\{ y_1^{(k)}, \ldots, y_{I_k}^{(k)} \right\} \,, \tag{3.8}$$

which contains the I_k items with the greatest target value among those with control class k, for all classes $k = 1, \ldots, K$.

Proof. For readers familiar with **matroids** we provide a conciser proof below.

Since all sets $R \in \mathcal{R}_{\boldsymbol{I}}(Q)$ have a constant coverage $|R| = \sum_{k=1}^{K} I_k$, maximising the objective value is equivalent to maximising the central tendency factor ${f_{\text{t}}}^{Q}$. We will show that $R^*_{\boldsymbol{I}}$ attains the greatest f_{t} value over $\mathcal{R}_{\boldsymbol{I}}(Q)$ by contradiction.

Assume there is a refinement $R' \in \mathcal{R}_{\boldsymbol{I}}$ with $R' \neq R^*_{\boldsymbol{I}}$ and $f_{\text{t}}(R') > f_{\text{t}}(R^*_{\boldsymbol{I}})$. Since $R^*_{\boldsymbol{I}}$ contains the items with maximum y value for each class, there is at least one sequence of refinements $(R^{(0)}, \ldots, R^{(T)})$, starting with $R^{(0)} \coloneqq R^*_{\boldsymbol{I}}$ and ending at $R^{(T)} \coloneqq R'$, so that at each index τ we exchange a single element between R^{τ} with another in $Q \setminus R^{\tau}$ of the same class, but a smaller target value. Formally, $R^{(\tau)} = \left(R^{(\tau-1)} \setminus \{\epsilon\}\right) \cup \{\epsilon'\}$, such that $c(\epsilon') = c(\epsilon)$ and $y(\epsilon') < y(\epsilon)$. This implies that, for each $\tau = 2, \ldots, T$ we get

$$\sum_{\epsilon \in R^{(\tau)}} y(\epsilon) - \sum_{\epsilon \in R^{(\tau-1)}} y(\epsilon) = y(\epsilon') - y(\epsilon) < 0 \,. \tag{3.9}$$

Dividing these sums with $\sum_{k=1}^{K} I_k$, turns them into means, and since f_t is increasing w.r.t. the target value mean, we have $f_\mathrm{t}(R^{(\tau)}) < f_\mathrm{t}(R^{(\tau-1)})$. By transitivity, it is $f_\mathrm{t}(R') < f_\mathrm{t}(R^*_{\boldsymbol{I}})$, contradicting the optimality of $f_\mathrm{t}(R')$, and concluding this proof.

Using the concept of matroids, an alternative, quicker proof can be given by observing that ${f_\mathrm{ct}}^{Q}(\boldsymbol{I})$ is a linear function to be optimised over the matroid $(E, \mathcal{R}_{\boldsymbol{I}}(Q))$, which has as ground set all entities and independent set all refinements with class count $\boldsymbol{I}$. Then, the optimal value is attained by greedy optimisation, which amounts to selecting the optimiser $R^*_{\boldsymbol{I}}$. □

As a result, we can express the target value mean of the optimiser $R^*_{\boldsymbol{I}}$ as $\mathrm{mean}(R^*_{\boldsymbol{I}}) = \sum_{k=1}^{K} \sum_{\mathrm{i}=1}^{I_k} y_\mathrm{i}^{(k)} / \|\boldsymbol{I}\|_1$, where $\|\boldsymbol{I}\|_1 = \sum_{k=1}^{K} I_k$ is the cardinality of each refinement in $\mathcal{R}_{\boldsymbol{I}}$. Now, for any given subset Q, the impact function f_ct of Eq. (2.9), when applied to any refinement $R \in \mathcal{R}_{\boldsymbol{I}}$ can be computed using the CCS representation of its equivalence class as

$$ {f_\mathrm{ct}}^{Q}(\boldsymbol{I}) \coloneqq \alpha_\mathrm{t} \sum_{k=1}^{K} \sum_{\mathrm{i}=1}^{I_k} y_\mathrm{i}^{(k)} - \alpha_\mathrm{c} \|\boldsymbol{I}\|_1 \,, \tag{3.10} $$

where

$$ \alpha_\mathrm{t} = \frac{1}{\nu} > 0 \,, \; \alpha_\mathrm{c} = \frac{\bar{y}_E}{\nu} \,, \text{ and } \; \nu = |E| \Big(\max_{\epsilon \in E} y(\epsilon) - \bar{y}_E \Big) \,. \tag{3.11} $$

Since the representativeness factor $f_\mathrm{r}(Q)$ depends only on the class counts of Q, it remains constant over $\mathcal{R}_{\boldsymbol{I}}$ and does not affect the maximiser. Therefore, the transformed controlled impact function can be written as

$$ {f_\mathrm{cif}}^{Q}(\boldsymbol{I}) \coloneqq {f_\mathrm{ct}}^{Q}(\boldsymbol{I})^{1-\gamma} \cdot {f_\mathrm{r}}^{Q}(\boldsymbol{I})^{\gamma} \qquad \gamma \in [0, 1) \;\; . \tag{3.12} $$

Notice that the value ${f_\mathrm{ct}}^{Q}(\boldsymbol{I})$ can be computed in constant time for any point $\boldsymbol{I} \in \mathcal{I}(Q)$, after a pre-processing step of linear time. Indeed, assuming the items of a candidate subgroup are in decreasing order of target values, we can achieve this by first passing through the values and creating K cumulative sums of target values, one for each class; after this step, the value ${f_\mathrm{t}}^{Q}(\boldsymbol{I})$ can be easily retrieved as the sum of indices I_k within the cumulative sum for each class k, appropriately scaled. The controlled impact function ${f_\mathrm{cif}}^{Q}(\boldsymbol{I})$ can be computed with trivial extra work to compute ${f_\mathrm{r}}^{Q}(\boldsymbol{I})$.

Therefore, this transform can be used in a straightforward way to derive an algorithm to compute the tight optimistic estimator in $O(n^K)$ time. However, a practical algorithm can benefit from further improvement, achieved in the next section for a widespread special case of a population.

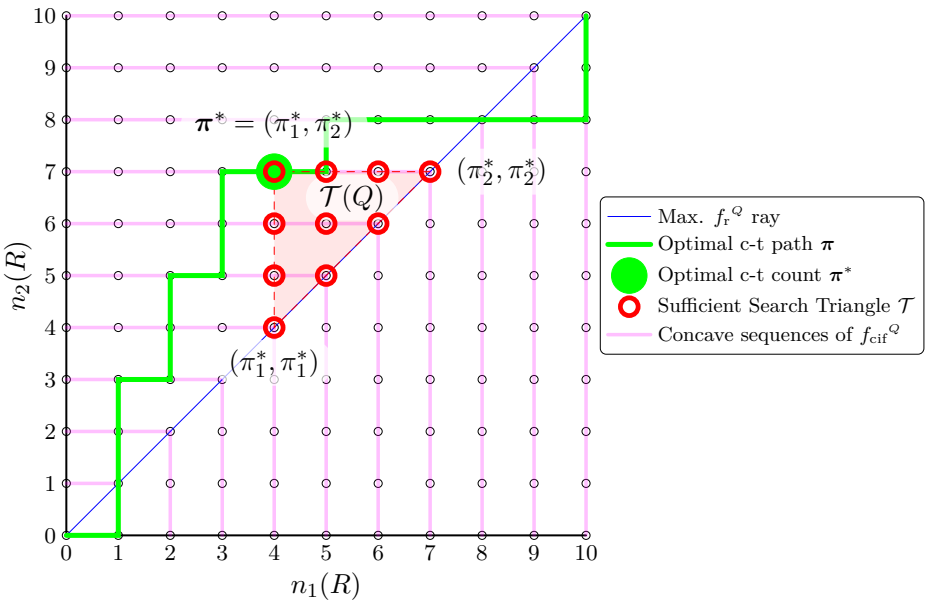


Figure 3.3: The sufficient search triangle $\mathcal{T}$ (red circles) in the CCS $\mathcal{I}$, and optimal point $\boldsymbol{\pi}^*$ along the c-t path (crooked line), which defines the 3 vertices of $\mathcal{T}(Q)$. We seek $\hat{f}_{\text{cif}} = \max f_{\text{cif}}{}^Q$, which lies on $\mathcal{T}(Q)$, and ternary search finds it efficiently along the concave sequences of $f_{\text{cif}}{}^Q$ (vertical/horizontal lines).

3.1.3 A Linearithmic Algorithm for Balanced Binary Controls

We now present a linearithmic algorithm able to compute the tight optimistic estimator of the controlled impact function of Eq. (3.1) for the widespread case of a population E with balanced binary control classes, i.e., $c\colon E \to \{1, 2\}$ with $n_1(E) = n_2(E)$.

The rest of the analysis can be summarised in two key steps. First, we show that there is a sub-region of $\mathcal{I}(Q)$ where $f_{\text{cif}}{}^Q(\boldsymbol{I})$ attains its maximum and then we present an efficient algorithm to search within this sub-region.

For this purpose we study the two factors, $f_{\text{ct}}{}^Q$ and $f_{\text{r}}{}^Q$ within the CCS. Both these factors form sequences that exhibit an appropriate notion of convexity for sequences, borrowed mutatis mutandis from Yucer [Yüc02]: a sequence $a : N \to \mathbb{R}$, with $N = \{0, \dots, n\}$ and $n \leq \infty$, is called a **convex sequence** over N, if for all $x, y \in N$ and each $\lambda \in (0, 1)$

$$\lambda a^{(x)} + (1 - \lambda)a^{(y)} \geq \min_{u \in \lfloor z \rfloor} a^{(u)}\,, \qquad z = \lambda x + (1 - \lambda)y\,. \tag{3.13}$$

Further, we call a a **concave sequence** if $-a$ is convex.

We now study the ${f_{\mathrm{ct}}}^Q$ values, as $|Q| = I_1 + I_2$ increases.

Definition 1 (Optimal c-t Path on $\mathcal{I}(Q)$)**.** *Let $\boldsymbol{\pi}^{(\mu)} \in \mathcal{I}(Q)$ be the maximiser of the ${f_{ct}}^Q$ value among all points in the CCS with a fixed sum μ*

$$\boldsymbol{\pi}^{(\mu)} \coloneqq \underset{\boldsymbol{I}\in\mathcal{I},\ \|\boldsymbol{I}\|_1=\mu}{\arg\max}\ {f_{ct}}^Q(\boldsymbol{I})\ , \qquad 0 \le \mu \le |Q|\ . \tag{3.14}$$

*We refer to the optimal point sequence $\boldsymbol{\pi} = (\boldsymbol{\pi}^{(0)}, \dots, \boldsymbol{\pi}^{(|Q|)})$ as the **optimal c-t path**.*

The optimal c-t path exhibits useful properties, discussed in the following lemma.

Lemma 3.2 (Optimal c-t path)**.** *Let $\boldsymbol{e}_1 = (1,0)^T$ and $\boldsymbol{e}_2 = (0,1)^T$ be the standard basis vectors of $\mathbb{R}^2$.*

i) Then the μ-th element of the optimal c-t path is the class count of the first μ elements of E; formally,

$$\boldsymbol{\pi}^{(\mu)} = \sum_{i=1}^{\mu} \boldsymbol{e}_{c_i} \qquad 0 < \mu \le |Q| \qquad \text{and} \qquad \boldsymbol{\pi}^{(\mu)} = \mathbf{0}\ . \tag{3.15}$$

ii) Moreover, the sequence ${f_{ct}}^Q \circ \boldsymbol{\pi}$, with elements the ${f_{ct}}^Q$ values computed along the c-t path $\boldsymbol{\pi}$, is a concave sequence.

— For the proof see Appendix A.1.

Lemma 3.2 allows for an $O(\log n)$ algorithm to find the **optimal c-t point** $\boldsymbol{\pi}^* \coloneqq \arg\max_{\boldsymbol{I}\in\mathcal{I}(Q)} {f_{\mathrm{ct}}}^Q(\boldsymbol{I})$, as we call the point f the CCS that maximises the ${f_{\mathrm{ct}}}^Q$ value. Indeed,

$$f_{\mathrm{ct}}(\boldsymbol{\pi}^*) = \max_{0\le\mu\le|Q|} \max_{\boldsymbol{I}\in\mathcal{I}(Q),\,|\boldsymbol{I}|=\mu} f_{\mathrm{ct}}(\boldsymbol{I}) = \max_{0\le\mu\le|Q|} f_{\mathrm{ct}}(\boldsymbol{\pi}^{(\mu)})\ , \tag{3.16}$$

where the last maximum runs over the ${f_{\mathrm{ct}}}^Q$ values of the optimal path sequence. Due to the concavity of this sequence, its maximum can be computed in $O(\log n)$ time, using for instance the ternary search algorithm.

We now study the representativeness factor $f_{\mathrm{r}}(Q)$, whose transform on the CCS for balanced binary controls becomes

$${f_{\mathrm{r}}}^Q(\boldsymbol{I}) \coloneqq 1 - \left|1 - \frac{2I_1}{I_1 + I_2}\right| = 1 - \left|1 - \frac{2I_2}{I_1 + I_2}\right|\ . \tag{3.17}$$

We observe that the subgroups $R \in Q$ that maximise this factor must have the same control class distribution as the population. Therefore, these

subgroups must have an equal control class count $n_1(R) = n_2(R)$, and thus belong to those equi-count refinement sets $\mathcal{R}_{\boldsymbol{I}}$, for which $I_1 = I_2$. These, in turn, lie on the so-called **maximum $f_{\mathrm{r}}{}^{Q}$ ray** $\boldsymbol{I} = (a, a)^T$, $a \geq 0$. As an example, we depict in Fig. 3.2b the maximum $f_{\mathrm{r}}{}^{Q}$ ray for the toy dataset of Fig. 3.2a; for balanced classes, a more complete example appears in Fig. 3.3.

We now state a key theoretical proposition of this section, showing that it suffices to search for the optimal solution on a specific triangle of the CCS.

Proposition 3.3 (Sufficient Search Triangle)**.** *The maximum of the controlled impact function $f_{cif}{}^{Q}$ is attained at a point which lies in the (filled) triangle $\mathcal{T}(Q) \coloneqq \{(\pi_1^*, \pi_1^*), (\pi_2^*, \pi_2^*), \boldsymbol{\pi}^*\}$, with vertices the optimal c-t point $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*)$ and its horizontal and vertical projections onto the maximum $f_r{}^{Q}$ ray. We call this region the* ***sufficient search triangle****.*

— For the proof see Appendix A.1.

The sufficiency of the SST reduces the search space by at least half, which happens in the worst case scenario that the optimal c-t point $\boldsymbol{\pi}^*$ is on the North-West or South-East points. More importantly, we can efficiently optimise $f_{\mathrm{cif}}{}^{Q}$ along specific directions within this region.

We now describe these directions. For each ordinate $\mathrm{i}_2 \in 0, \ldots, n_2(Q)$ let the (West-to-East) **horizontal sequence** be

$$\boldsymbol{h}_{\mathrm{i}_2} \coloneqq (\boldsymbol{h}_{\mathrm{i}_2}^{(0)}, \ldots, \boldsymbol{h}_{\mathrm{i}_2}^{(n_1(Q))}) = \big((0, \mathrm{i}_2), \ldots, (n_1(Q), \mathrm{i}_2)\big) \,. \tag{3.18}$$

Similarly, for each abscissa $\mathrm{i}_1 \in 0, \ldots, n_1(Q)$ we define the (South-to-North) **vertical sequence**

$$\boldsymbol{v}_{\mathrm{i}_1} \coloneqq (\boldsymbol{v}_{\mathrm{i}_1}^{(0)}, \ldots, \boldsymbol{v}_{\mathrm{i}_1}^{(n_2(Q))}) = \big((\mathrm{i}_1, 0), \ldots, (\mathrm{i}_1, n_2(Q))\big) \,. \tag{3.19}$$

When the transformed controlled impact function $f_{\mathrm{cif}}{}^{Q}(\boldsymbol{I})$ is computed along the elements of certain of those sequences, it forms concave sequences, as we show below, with the direct implication that the maximal value of f_{cif} along these sequences can be computed in $O(\log n)$.

Proposition 3.4 (Concavity of $f_{\mathrm{cif}}{}^{Q}$ along sequence)**.** *Consider the values of the controlled impact function $f_{cif}{}^{Q}$ as they are computed along a horizontal sequence $\boldsymbol{h}_{\mathrm{i}_2}$; these form the sequence $(f_{cif}{}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})(\mu)$, which for $\mu \leq \mathrm{i}_2$ is a concave sequence preceding the maximum $f_r{}^{Q}$ ray. Similarly, $(f_{cif}{}^{Q} \circ \boldsymbol{v}_{\mathrm{i}_1})(\mu)$ is a concave sequence for $\mu \leq \mathrm{i}_1$.*

— For the proof see Appendix A.1.

Observing the example of the concave sequences of $f_{\mathrm{cif}}{}^{Q}$ along the horizontal and vertical directions shown in Fig. 3.3, we notice that we can cover the

entire SST with an appropriate selection of these concave sequences. This allows for an efficient optimisation procedure requiring $O(n \log n)$ time, which is described in Algorithm 2 and operates as follows.

Algorithm 2: RAWR

Input: Population E (sorted w.r.t y, descending)
Input: Subgroup Q
Output: Tight optimistic estimator $\hat{f}_{\text{cif}}$ of Eq. (3.1)

1 $\boldsymbol{\pi}^* \leftarrow$ **TernarySearch**(on ${f_{\text{ct}}}^Q \circ \boldsymbol{\pi}$ from 1 to $|Q|$);
2 $\text{i}_{\text{beg}} \leftarrow \min\{\pi_1^*, \pi_2^*\}$;
3 **if** $\pi_1^* < \pi_2^*$ **then**
4 $\text{i}_{\text{end}} \leftarrow \min\{\pi_2^*, n_2(Q)\}$;
5 **for** i from i_{beg} to i_{end} **do**
6 $\phi \leftarrow$ **TernarySearch**(on ${f_{\text{cif}}}^Q \circ \boldsymbol{h}_{\text{i}}$ from i_{beg} to i_{end});
7 $\hat{f}_{\text{cif}} \leftarrow \max\{\hat{f}_{\text{cif}}, \phi\}$;
8 **else**
9 $\text{i}_{\text{end}} \leftarrow \min\{\pi_1^*, n_1(Q)\}$;
10 **for** i from i_{beg} to i_{end} **do**
11 $\phi \leftarrow$ **TernarySearch**(on ${f_{\text{cif}}}^Q \circ \boldsymbol{v}_{\text{i}}$ from i_{beg} to i_{end});
12 $\hat{f}_{\text{cif}} \leftarrow \max\{\hat{f}_{\text{cif}}, \phi\}$;
13 **return** $\hat{f}_{\text{cif}}$;

First, the optimal c-t point $\boldsymbol{\pi}^*$ is computed in $O(\log n)$ time, along the concave sequence $\boldsymbol{\pi}$ (line 1); this point locates the SST. If $\boldsymbol{\pi}^*$ lies above the maximum ${f_{\text{r}}}^Q$ ray (line 3-7), the points of $\mathcal{T}(Q)$ lie along horizontal sub-sequences preceding the maximum ${f_{\text{r}}}^Q$ ray; the ${f_{\text{cif}}}^Q$ values along each of them form a concave sequence, whose maximum can be found in $O(\log n)$ (ln. 6). There are at most $n_2(Q)$ such directions in $\mathcal{T}(Q)$, and they can all be scanned (ln. 5-7) in a total of $O(n \log n)$ time. Similarly, when $\boldsymbol{\pi}^*$ lies below the maximum ${f_{\text{r}}}^Q$ line (ln. 8), we optimise along the vertical sub-sequences (ln. 9-12).

3.2 Related Work

Although subgroup discovery is a well-studied problem in general, (see Section 1.2 for an overview), we are the first to study the notion of representative subgroups.

When it comes to rejecting already known trends, a general approach in

pattern mining has been to discover patterns that are surprising given background knowledge, for example with regard to permutation testing [GMM+07; HOV+09], or a maximum entropy distribution [Tat08; KB10; KVD11; MVT12]. While seemingly related, representativeness is not guaranteed by unexpectedness: adding a pattern X to our background knowledge does not ensure that, relative to X, the now-most-surprising pattern will be representative with regard to either pattern X, or to the whole population.

Another seemingly obvious relation that turns out to be much more subtle is that to fairness in classification. A truly representative pattern implies statistical parity with regard to the control variable, although it is worth noting that both Dwork et al. [DHP+12] and Kleinberg et al. [KMR16] explicitly mention that statistical parity should not be equated with fairness, as it can potentially be "blatantly unfair" on an individual level [ZWS+13; KMR16].

In recent work, Feldman et al. [FFM+15] studied the notion of "disparate impact"—a legal term that says that the probability ratio of treatment (e.g., job offer) for the different groups must be at least 0.8—and proposed as a general technique to remove disparate impact via data pre-processing. In other words, unlike our approach, the global distribution is changed to de-correlate sensitive and target attributes. Related as it may be, their work clearly fails to extend to local pattern mining, as in the latter, it does not suffice to model the global distribution.

Perhaps closest to our approach is the line of work by Calders et al. [CŽ13], who studied the goal of achieving statistical parity in classification with different methods, including naive bayes [CV10] and decision trees [KCP10]. Kamishima et al. [KAA+12] consider a form of fairness that is related to statistical parity, although implicitly: during logistic regression a regularisation term is used, measuring the KL divergence between sensitive attribute and prediction. Although related, it is unclear whether these methods can be utilised in the highly demanding branch-and-bound search, typically able to optimise over exponentially-sized discrete spaces of arbitrary subgroup descriptor languages.

Closest in terms of pattern mining, but relatively distant with regard to statistical parity, is the work by Pedreschi et al. [PRT08] on discrimination-aware pattern mining. Instead of subgroups, the authors focus on mining association rules that may only include a sensitive item if this does not improve the confidence of the rule by more than α.

	Dataset	Target y	Control c	α	$\lvert V\rvert$	$\lvert E\rvert$
Qualitative	baseball	Salary	Fr.Ag.Elig.	1.0	16	268
	gold	Evdw-Evdw0	Odd	1.0	19	12200
	homicide	Victims	PerpRace	1.0	10	47236
	students	G3	failures	0.5	31	366
	wine	quality	colour	1.0	12	3198
Quantitative	abalone	Rings	Height	1.0	8	4144
	ailerons	Sa	RollRate	1.0	5	7108
	airfoil	NoiseLevel	Displacement	1.0	5	1480
	autompg	Mpg	Cylinders	1.0	8	380
	bike	registered	atemp	1.0	13	730
	california	Med.Value	Latitude	0.5	8	20502
	compactiv	usr	freeswap	0.7	21	8192
	concrete	Strength	Age	1.0	8	562
	elevators	Goal	DiffRollRate	0.3	18	16020
	forestfires	Area	Month	0.6	12	394
	house	Price	P14p9	0.3	16	22784
	mortgage	30YRate	Mat.Rate3Y	0.8	15	1044
	mv	Y	X6	0.9	10	40768
	pole	Output	Att0	0.3	26	14586
	puma32h	thetadd6	theta5	0.7	32	8192
	stock	Company10	Company4	1.0	9	950
	treasury	X1Rate	CMat.Rate3Y	0.4	15	1044
	wankara	AvgTmp	MaxTemp	0.6	9	318
	wizmir	AvgTmp	MaxTemp	0.5	9	1458

Table 3.1: Used datasets, for qualitative (top) and quantitative (bottom) analysis. Listed are the number of attributes $|V|$ and number of rows $|E|$, as well as running configurations: target and control variables, and approximation factor α. The latter is decreased by 0.1 every time **b**inary **r**epresentativeness **ig**norant (BRIG) exceeds a timeout of 6 hours, or terminates due to exceeding 256GB of memory.

3.3 Experiments

In this section we evaluate our extended impact function f_{cif}, as well as RAWR, our implementation of its tight optimistic estimator. We provide qualitative and quantitative demonstrations of superior representativeness in the discovered subgroups, and we also report runtime measurements on a variety of datasets.

For the sake of both these tasks we implemented[1] both RAWR and the non-tight, representativeness oblivious **b**inary **r**epresentativeness **ig**norant (BRIG), which we use as a baseline. We then run the branch-and-bound algorithm equipped with each optimistic estimator.

3.3.1 Mining Representative Results

We now assess qualitatively and quantitatively the representativeness of the discovered subgroups, for different values of the γ parameter. We first study 5 datasets retrieved from the UCI ML repository[DG17] and the Murder Accountability Project[2], which contain intuitively interpretable controls (Table 3.1, top). To rule out the effect of unbalanced classes, and for our algorithm to be applicable, we stratify the datasets over the control classes. We then perform subgroup discovery as a multi-objective optimisation (MOO) problem (see Section 2.5) with the scores f_{r} and f_{ct}, as we sweep the γ parameter (Fig. 3.4).

Obviously, a value of $\gamma = 0$ corresponds to the representativeness-oblivious impact function f_{ct} of Eq. (2.9). Depending on the dataset and choice of y and c, the discovered subgroups for this case may be representative, although this is not guaranteed. However, as the γ parameter increases, the added f_{r} factor comes in effect, yielding consistently more representative subgroups [top]. As expected, the f_{ct} score may drop, demonstrating that γ controls the trade-off between the two factors [bottom]. At the same time, it is guaranteed that no score can be increased without the decrease of the other, by choosing a subgroup other than the discovered.

We next delve into the subgroups discovered in selected datasets. We first focus on the `Homicide` dataset, which tracks homicide cases, matched with background data on perpetrators and their victims, alongside the number of victims per case. We use the latter as a target variable to measure violence and seek to gain insight on attributes leading to increased violence, as captured by binary control variables. For each studied variable, we stratify

[1] Our source code is available within the realKD tool `bitbucket.org/realKD/`.
[2] Available at `http://www.murderdata.org/`.

the dataset and report the discovered subgroups (Table 4.3), for increasing γ parameter.

We first consider the effect of the `Perpetrator Sex`. The subgroup discovered without the f_{r} term rediscovers the unsurprising fact that males are more violent than females. To uncover further potentially underlying factors, we use the `Perpetrator Sex` as a control variable and perform subgroup discovery using the controlled impact function. As γ parameter increases, the discovered subgroups hold for both male and female perpetrators, leading to the discovery that "*Caucasian victims attract more violence*", and further that *no sex is more violent when it comes to older victims.*

Controlling for a variable can also serve as a way to disprove the contribution of a sensitive variable toward the effect. Consider for instance the sensitive variable `Perpetrator Race`, appropriately controlling for which we find that "*both races are (equally) more violent when murdering younger females*". Further observing the behaviour of the algorithm as we lift the constraint of representability by lowering the γ parameter, we see that the discovered sub-population remains virtually the same. This very fact provides substantial evidence against a hypothesis that one race is more prone to violence, at least when it comes to this subgroup. In other words, if the race of the perpetrator was indeed a factor that affected this case, then the algorithm would be able to increase the objective by simply restricting this subgroup with a single added predicate that constrained the sub-population within the specific race. From a statistical perspective, subgroup discovery can find sub-populations, within which it becomes interesting to further test for the dependence between the target and control variables—say if we additionally assume a model for higher average violence by some specific race. Overall, in this case, the fact that no better subgroup appears when we lift the constraint for representability provides evidence that "*no race is more violent when murdering young women*".

In another example, we study an application for fair subgroup discovery. Consider that a baseball team decides to increase its players salary and seeks to find the factors that lead to higher income drawing experience from other team managements. At the same time, the raise should not be unfavourable to players who are contractually bound to one particular team, in contrast to the Free Agent eligible players, which might earn more lest they leave the team. Using the `FreeAgencyEligibility` variable as control, more objective criteria describing high salaries are discovered.

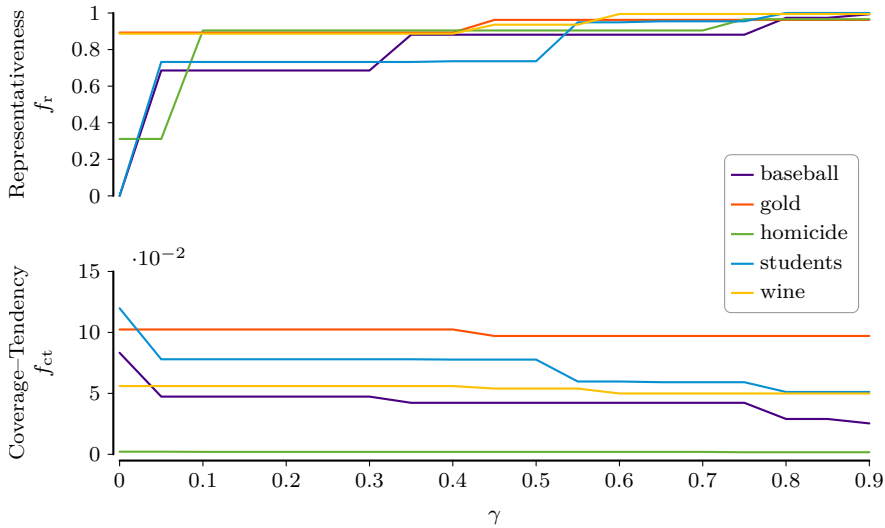


Figure 3.4: Scores of subgroups discovered in the qualitative datasets (Table 3.1, top). Tuning γ effectively controls the trade-off between representativeness and coverage–tendency.

3.3.2 Performance of our Tight Optimistic Estimator

We now evaluate the performance of the RAWR implementation. To sample a broad variety of datasets, we used all of the regression datasets from the KEEL database [AFL+11] with a number of variables $8 \leq |\boldsymbol{V}| \leq 40$ (Table 3.1, bottom). As a target variable we used the designated regression variable. To emulate a purported scenario of controlling for the main data trend, we use as control the first variable that appears in the subgroup descriptor discovered for $\gamma = 0$; if this variable is real we discretise it around the median. Next, we stratify the dataset on the control variable. We start with an approximation factor of $\alpha = 1$, corresponding to exact computation; when all BRIG invocations for a dataset fail, due to either a runtime of more than 6 hours or exceeding 256GB of available memory, we decrease α by 0.1 and repeat.

We assess the performance of our algorithm w.r.t. the number of searched nodes during each invocation and also the needed runtime (Fig. 3.5); we set $\gamma = 0.6$, corresponding to a reasonably practical scenario. As our proposed optimistic estimator is tighter, it is yielding a significantly better pruning performance. What is more, our implementation seems to make use of the better pruning achieved, in order to attain running times that are comparable to those of BRIG, or in some cases up to 4 orders of magnitude

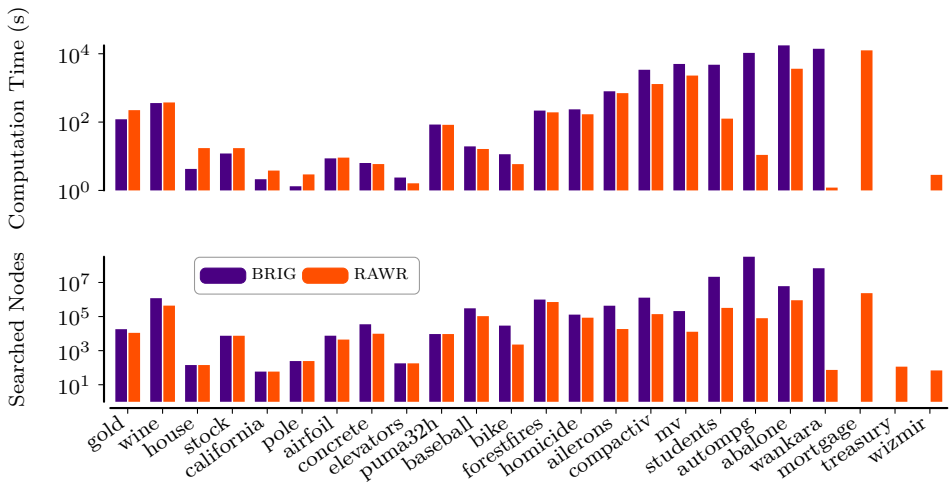


Figure 3.5 [Lower is better]: Performance comparison of RAWR (solid) and BRIG (dashed), for runtime [top] and searched nodes [bottom] for $\gamma = 0.6$. The datasets (x axis) are sorted in increasing time difference. On the 3 last datasets BRIG exceeds the 256GB of our available memory, which prevents computing the result.

better. Further numerical results are reported in Table 3.2, for a set of sensible weights $\gamma \in \{0.4, 0.5, 0.6\}$. These show a superiority of our estimator especially for higher γ values, where BRIG is less tight.

Furthermore, note that the number of nodes is a key factor contributing to the memory requirement of the branch and bound algorithm. As such, even for dataset on which the computation time of these implementations might be comparable, it is sometimes the case that the decreased number of nodes is enabling the computation using RAWR, where otherwise BRIG would exceed available memory, e.g., in the last 3 datasets of Table 3.2.

3.4 Discussion

Our introduced method guarantees the optimality of the results given the specified parameter, while optionally enabling a faster computation by relaxing the optimality guarantee.

The sole parameter γ of our method remains intuitive in its interpretation (see Section 2.5) and possibly in its selection, regardless of the input, with the zero value corresponding to a vanishing effect of our extension and a high value an increased weight of it. Nonetheless, not every dataset is equally

sensitive to the intermediate values and the researcher is still required to make educated guesses based either on expert knowledge or a trial-and-error scheme.

As a downside, our implementation of the tight estimator for our objective function requires a dataset with balanced, binary classes. Nevertheless, the case of binary classes is amongst the most widespread, and obtaining a balanced dataset can be solved by using as a workaround an appropriate stratification as a pre-processing step.

We leave as future work the investigation on how to obtain a log-linear optimistic estimator in conditions other than balanced binary classes. We also note that this chapter already provides an algorithm to compute the tight optimistic estimator also for any case beyond binary classes, albeit with polynomial computational complexity; this is still a considerable improvement on the otherwise exponential complexity of naïve algorithms.

3.5 Conclusion

In this chapter we introduced the problem of representative subgroup discovery, where our goal is to discover subgroups that are exceptional with regard to the target variable, yet at the same time have statistical parity with respect to the control variable. We show how we can achieve this by extending the typically used impact function to incorporate a tuneable representativeness factor. We propose a tight optimistic estimator for the newly representative aware impact function, and give an efficient algorithm to compute it in $O(n \log n)$ time. Our experiments show that our proposed method may lead up to orders of magnitude fewer node expansions, compared to the representative ignorant estimator, the direct effect of which is a speedup of similar magnitude.

	$\gamma = 0.4$		$\gamma = 0.5$		$\gamma = 0.6$	
Dataset	**rawr**	**brig**	**rawr**	**brig**	**rawr**	**brig**
gold	172	**101**	210	**99**	224	**121**
wine	296	**267**	349	**305**	375	**360**
house	14	**5**	13	**5**	17	**4**
stock	15	**8**	16	**10**	17	**12**
california	3	**2**	4	**2**	4	**2**
pole	4	**2**	3	**2**	3	**1**
airfoil	**9**	9	**7**	8	9	**9**
concrete	**4**	6	**5**	5	**6**	6
elevators	3	**3**	3	**3**	**2**	2
puma32h	**78**	84	**82**	83	**83**	85
baseball	**20**	22	**17**	21	**16**	19
bike	**5**	10	**5**	12	**6**	11
forestfires	**184**	272	**178**	186	**193**	217
homicide	**154**	219	**171**	247	**169**	236
ailerons	**206**	317	**297**	486	**703**	797
compactiv	504	**450**	442	**349**	**1299**	3397
mv	**3877**	6837	**3243**	5273	**2300**	5026
students	**19**	175	**63**	2638	**126**	4768
autompg	**6**	3229	**6**	5577	**11**	10591
abalone	**869**	1307	**1883**	3876	**3639**	17575
wankara	**1**	116	**1**	2543	**1**	13977
mortgage	**56**	∞	**198**	∞	**12568**	∞
treasury	**1**	**1**	**1**	**1**	**1**	∞
wizmir	5	**3**	**2**	1648	**3**	∞

Table 3.2 [Lower (bold) is better]: Runtime comparison of RAWR and BRIG over different γ parameters for all datasets, sorted in increasing time difference. Using BRIG on the last 3 datasets exceeds our 256GB of memory, so results are not available.

	γ	Subgroup describing Q	$f_{\mathrm{r}}(Q)$	$f_{\mathrm{ct}}(Q)$
homicide		Control: `Perpetrator Sex`		
	$(0, 0.09]$	Crime=Murder, Vict.=White, Perp=♂	0.00	0.002
	$(0.09, 0.75]$	Vict.=White	0.89	0.002
	$[0.75, \dots)$	Vict.Age= ¬V.Lo, Vict.=White	0.99	0.001
		Control: `Perpetrator Race`		
	$(0, 0.6]$	Crime=Murder, Vict=♀, Perp.= ¬V.Old	0.90	0.002
	$[0.6, \dots)$	Crime=Murder, Vict=♀, Perp.= ¬Old	0.98	0.002
		Control: `Vict. Sex`		
	$(0, 0.09]$	Vict=♀, Perp=♂, Perp.Age= ¬V.Hi	0.00	0.003
	$(0.09, 0.47]$	Vict.=White, Perp=♂, Perp.Age= ¬V.Hi	0.90	0.002
	$(0.47, 0.8]$	Vict.=White, Perp=♂, Perp.Age= ¬HI	0.98	0.002
	$[0.8, \dots)$	Vict.=White, Perp.Age= ¬HI	0.99	0.002
baseball		Control: `Free Agent Eligibility`		
	$(0, 0.09]$	OnBase= ¬V.Lo, F.A.Eligible= ✓	0.00	0.083
	$(0.09, 0.33]$	OnBase=HI	0.69	0.047
	$(0.33, 0.8]$	Batting= ¬V.Lo, OnBase= ¬Lo	0.88	0.042
	$[0.8, \dots)$	Batting=¬V.Lo, OnBase= ¬Lo, Fr.Ag.= ✗	0.97	0.029

Table 3.3: Discovered subgroups for a varying γ parameter, for the datasets `homicide` (above) and `baseball` (below). Increasing γ produces more representative subgroups.

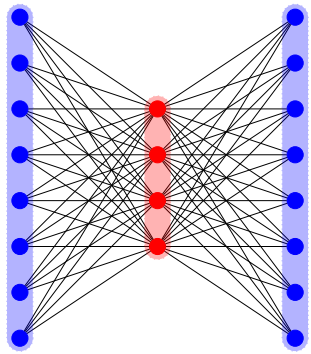
4 Robustly Connected Subgroups

Everything is connected.

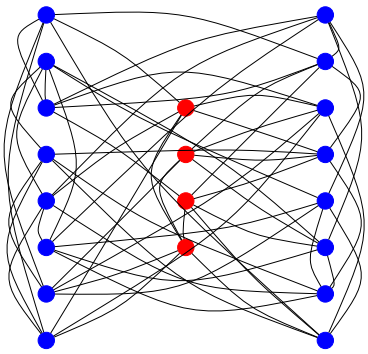
(Alan Moore, V for Vendetta)

Up until now, in the subgroup discovery methods we studied we assumed that all entities of the dataset were independent with one another and identically distributed (i.i.d.). Despite its convenient, however, in several important applications this is an assumption that we cannot afford to make, since there is additional structural information between the entities that is crucial for an interesting result. Of particular interest is a versatile kind of structural information that can be captured in the form of pair-wise relations between entities. These pairwise relations can, in turn, be modelled as a graph, whose vertices represent entities and whose edges represent the existence of a pair-wise relation between the corresponding entities.

This natural way to represent pair-wise relationships between entities arises naturally in numerous applications, for instance in the study of resource allocation, social interactions, or knowledge representations, just to name a few. In such applications a particularly useful information is conveyed in the subgraphs that are exceptionally well-connected, and discovering such subgraphs has therefore attracted a lot of attention across different scientific domains. A large part of this attention has been focused on discovering dense subgraphs, and numerous methods are available that can report these subgraphs by simply listing them a bag-of-vertices, that is a list of the vertex identifiers that partake in the dense subgraph. From the perspective of this dissertation, it should already be evident that such methods suffer a major drawback: that the do not convey intuitive information to a human audience of what these vertices have in common—they lack an intelligible description. As we established, without the additional constraint that the discovered subgraphs have such a description, their vertices can be arbitrarily chosen to maximise the selected measure. Apart from being difficult to interpret, however, the resulting vertex subsets are possibly not even interesting to begin with. We therefore approach this task from the perspective of subgroup discovery, thus requiring that we can only select well-connected subgraphs that are also describable. For instance, the description of the most well-



(a) complete bipartite graph
edge/vertex ratio: 3.2—coreness: 4.



(b) 6-regular graph (also a 6-core)
edge/vertex ratio: 3—coreness: 6.

Connectedness Measure	Graph		
	bipartite (a)		**6-regular (b)**
intuitively connected	*weakly*	$<$	*robustly*
removed vertices until edgeless	4	$\ll$	19
edge-to-vertex ratio	3.2	$>$ ✗	3
average coreness	4	$\ll$ ✓	6

(c) Comparison of connectedness metrics.

Figure 4.1 [Edge-to-vertex ratio vs. robust connectedness]: Out of these two graphs with the same number of vertices, the bipartite graph a has more edges then the 6-regular graph b, and is therefore more densely connected in terms of the edge-to-vertex ratio. Nevertheless, graph b is much more *robustly* connected than graph a, since we can make graph a *fully disconnected* by removing just its 4 central nodes, whereas to achieve the same for graph b we need to remove 19 vertices (!).

connected subgroup found within data borrowed from the Internet Movie Database (IMDB) reveals that "the mainstream movie crew members with lengthy experience have collaborated together more than usual in the movie industry".

Moreover, in this work we depart from the notion that subgraphs with high edge-to-vertex ratios are interesting per se, as is assumed in most established dense subgraph methods [ATM+12; PBV14; GGT16; PA18]. Despite its appeal at first glance, it is a rather naïve measure of whether vertices 'belong together', as it only considers numbers of edges rather than their structure.

As an example, let us consider the two toy graphs of Fig. 4.1, each of which has 20 vertices. The bipartite graph on the left has a high edge-to-vertex ratio, but is arguably not very robustly connected; that is, we can fully disconnect it by only removing the 4 central nodes. In contrast, the graph on the right has a lower edge to vertex ratio, but is robustly connected: to disconnect it, we would have to remove virtually all but one vertices, which amounts to 19 vertices. A structure like the leftmost graph can frequently arise when within a sub-population there are a entities that have a large outreach, even though the rest of the sub-population has a considerably lower standing—a very fitting example can be a group of social network members that follow a few influential figures in a social network. Although knowing of such a group of members can be useful in certain scenarios, the function of each is very different within the subgraph and; that is, apart from these few influential figures connecting them, they can otherwise have little in common, thereby not even forming a sub-population worth our while in the first place. On the contrary, the entities of the rightmost subgraph in Fig. 4.1 are all equally well-connected with each other and constitute a more coherent sub-population; this structure is likely to arise as the result of an important mechanism, which therefore signifies a subgroup worth discovering. As an example, we mention once more the list of movie crew members that have each collaborated tightly with one-another. Importantly, as we can also see in this example, such a sub-population is *robustly* connected, that is, it retains its characteristic structure even after removing quite a few of its members.

We hence study the problem of discovering *robustly* connected subgraphs that admit an intelligible description. We propose a score for the robust connectedness of subgraphs based on the notion of k-coreness, which underlies our intuition from Fig. 4.1, and use it within the framework of (multi-objective) subgroup discovery. Interestingly, we further establish our method within the plethora of works of the field, by demonstrating that, even though the existing description-agnostic methods can later be endowed with descriptions, these correspond to subsets of arbitrarily low quality in terms of the adopted measures. In this work we also develop a tight optimistic estimator that we use within our IDDFS framework (see Section 2.4.2) to efficiently find the optimal subgroup in the language of closed selectors $\mathcal{L}_{\text{cl}}$ (see Eq. (2.8)), thus giving rise to our RoSi method.

Extensive experiments on large and diverse real-world graphs show that RoSi performs very well in practice, discovering meaningful subgraphs while competing ones run out of time and memory. Finally, we also compare with state-of-the-art methods which do provide descriptions, and therefore

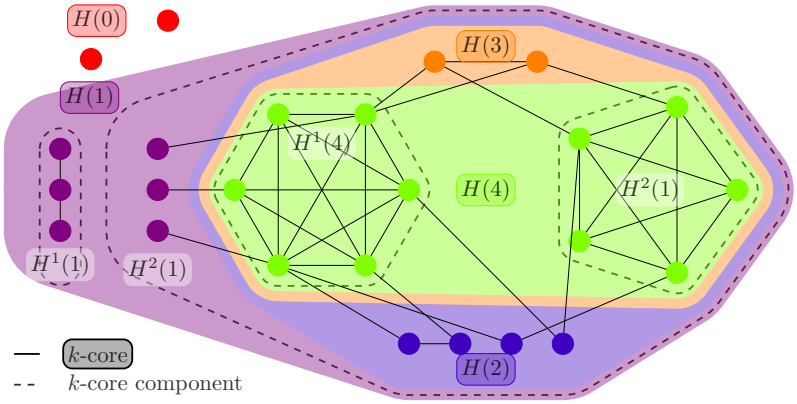


Figure 4.2 [Higher coreness coincides with higher density]: The core decomposition of a graph hierarchically groups its vertices into increasingly denser subgraphs. Here $H(k)$ denotes a k-core and $H^i(k)$ the i-th k-core component.

provide a qualitative comparison between their result of a subgroup with high edge-to-vertex ratio, against our novel measure of robust connectedness. These experiments also show that the above example is not esoteric: the densest subgraph that the recent method LDENSE [GGT16] discovers from `DBLP` is one with high average density but a robustness of 0, indicating that its robust connectedness is even lower than the robust connectedness in the entire graph (!).

4.1 Core Decomposition: k-Cores and their Coreness

Before we study the main measure we are going to use, we introduce the notion of k-cores [Bic10]: a useful measure of structural connectedness for each vertex that reveals important structural properties of each vertex [SEF16].

Assume a graph $G = (V, E)$ and a vertex subset $U \subseteq V$, whose connectivity we want so study. This vertex subset U defines on G the **induced subgraph**, i.e., the subgraph $G[U] \coloneqq (U, E(U))$, where $E(U) \coloneqq \{(v, u) \in E \mid u, v \in U\}$ is the set of all edges with end-points in U. For a vertex v, we denote by $N(v) \coloneqq \{u \in E \mid (u, v) \in E\}$ its **neighbours** in G and its **degree**, i.e., the number of its neighbours, by $\delta(v) \coloneqq |N(v)|$. When a quantity refers to the induced graph $G[U]$ we indicate the inducing vertex set as a subscript. For instance, $\delta_U(u)$ denotes the degree of vertex u in the induced graph $G[U]$.

A **k-core component** of a graph G is an (inclusion-wise) maximal connected subgraph of G whose vertices U have all a degree of at least $\delta_U(u) \geq k$. The subgraph comprising all k-core components of this graph is called its

k-core $H(k)$, and the **k-core vertices** $E(k)$ are the vertices of the graph's k-core. The last two definitions are then related as $H(k) \coloneqq G[E(k)]$.
The annotated k-cores of the example graph on Fig. 4.2 show that the k-cores are nested to form a hierarchy over the vertices. We also define the **k-shell** of G as the set of vertices that lie in the k-core but not in the $k+1$-core (same-coloured vertices in the figure). In this way, the k-shells define a partitioning over the vertices: the **core decomposition** of G, which assigns to each vertex v a **core number** (or **coreness**)

$$\kappa(v) \coloneqq \max\{k \mid v \in E(k)\} \ , \qquad \text{for } G, \text{ and} \tag{4.1}$$

$$\kappa_U(v) \coloneqq \max\{k \mid v \in E_U(k)\} \ , \qquad \text{for } G[U] \ , \tag{4.2}$$

equal to the greatest number k such that this vertex lies in the k-core of G, where $E_U(k)$ are the k-core vertices of $G[U]$. Note that by definition $G[V] \equiv G$, and hence $\kappa_E(v) \equiv \kappa_V(v)$. Finally, the graph **degeneracy** $K \coloneqq \max_{v \in E} \kappa(v)$ is the maximum coreness over all the vertices of the graph.

Note that graph coreness is related to various definitions of density [SEF16]: high coreness indicates better connectedness. For instance, the minimum coreness in a graph lower bounds the number of edges that have to be removed for the subgraph to become disconnected.

4.2 Measuring Robust Connectedness

We can now lay the basic notation for the study of connected entities within the subgroup discovery framework. We study sets of entities, for which we are given attribute values as well as structural information in the form of connections between them. Solely for this chapter, and to remain aligned with the established norms in graph-related context, we replace our notation for an entity to be that of V. Hence, we consider vertex-attributed (multi-)graphs $G = (E, E, X)$, where the vertices V correspond to entities and the edges E to connections between them. The set of vertex attributes $X \coloneqq \{x_1, \ldots, x_p\}$ comprises assignments $x_i : V \to \mathcal{X}_i$ from vertices to a continuous or categorical domain $\mathcal{X}_i$. These attributes can be used to simply describe subsets based on logical expressions of vertices $v \in V$ like $s(v) \equiv [\texttt{age}(v) \geq 18] \wedge [\texttt{sex}(v) = \text{`}\textit{female}\text{'}]$.

Our goal is to identify such logically described sets of vertices $U \subseteq V$ that are relatively large but also more robustly connected than G as a whole. That is, we aim to identify significant parts of the graph that stand out due to their connectedness. Note that size and connectedness are inversely

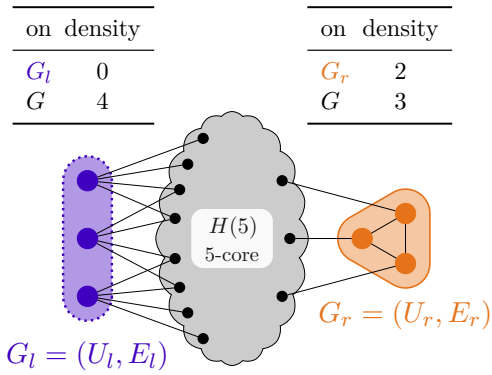


Figure 4.3: The average subgraph coreness $\bar{\kappa}_U \equiv \bar{\kappa}_U(U)$ may be misleadingly overestimated when it is computed on the whole graph $\bar{\kappa}_E(U)$. Here, subgraph G_r is denser than G_l with $\bar{\kappa}_{U_r} = 2 > 0 = \bar{\kappa}_{U_l}$. However, counting the edges of G, the subgraph densities falsely indicate the opposite relation. The same artefact appears to an even greater extreme for the naïve measure of edge-to-vertex ratio.

related: while it is easy to construct a small U with highly connected vertices, a large U must also include loosely connected ones. We hence maximise their (weighted) multiplicative trade-off, called **density impact**, defined as

$$f_{\mathrm{di}}(U;\gamma) \coloneqq f_{\mathrm{c}}(U)^{(1-\gamma)} f_{\mathrm{d}}(U)^{\gamma} \qquad \text{with } \gamma \in (0,1)\ , \tag{4.3}$$

where γ is a **trade-off parameter** that tunes the importance between the **coverage term** $f_{\mathrm{c}}(U) \coloneqq {}^{|U|}/_{|V|}$, i.e., the portion of the graph covered by the subset U, and the **density term** $f_{\mathrm{d}}(U)$, which increases as the vertices in U become more robustly connected. This objective function bears similarities to the geometrically weighted impact function of Eq. (2.10), where the exceptionality measure is replaced with a density-aware one. Therefore, it boasts similar advantages (see Section 2.3), and, importantly, also allows for a multi-objective optimisation approach for our task (see Section 2.5).

We can now employ the useful notion of coreness to formally define the very same measure of density that we encountered in Fig. 4.1 to quantify robust connectedness. Namely, we define the **average coreness** of G and $G[U]$, respectively, as the mean coreness of its vertices

$$\bar{\kappa}_V \coloneqq \frac{\kappa_V}{|V|} \text{ and } \bar{\kappa}_U \coloneqq \frac{\kappa_U}{|U|} \text{ with } \kappa_U \coloneqq \sum_{v \in U} \kappa_U(v) \text{ for } U \subseteq V. \tag{4.4}$$

We hence quantify the amount to which a vertex set U is more robustly connected than G by the **coreness density**

$$f_{\mathrm{d}}(U) \coloneqq \bar{\kappa}_U - \bar{\kappa}_V\ , \tag{4.5}$$

which completes the definition of the density term of Eq. (4.3).

4.3 Discovering Robust Describable Subgraphs

Our goal is hence to identify large and robustly connected vertex sets which have an intelligible description. Formally, we can specify our problem as finding the subgroup that maximises

$$s^* \in \arg\max_{s \in \mathcal{L}} f_{\mathrm{di}}(U;\gamma)(s)\,, \tag{4.6}$$

where $f_{\mathrm{di}}(U;\gamma)$ is the family of objective functions defined in Eq. (4.3). This problem can be optimised exactly using the general framework of our IDDFS algorithm, as presented in Section 2.4.2. This algorithm, in turn, requires an efficient optimistic estimator, which we derive below.

4.3.1 The Tight Optimistic Estimator

To derive an optimistic estimator for our objective function, we need to show that it satisfies the bound of Eq. (2.27). A first such bound can be derived by adapting ideas from rule mining (or subgroup discovery) on numerical unstructured data [GR09]. Here, each entity v has a real-valued *target attribute* y, and our aim is to find a describable subset $U \subseteq V$ in which the mean value of y is maximal. Using coreness as the target attribute, the formal objective in this task becomes a static version f^s_{di} of our f_{di}:

$$f^s_{\mathrm{di}}(U) = \frac{|U|}{|V|}\left[\sum_{u\in U} \bar{\kappa}(u) - \bar{\kappa}\right]\,, \tag{4.7}$$

where v^s_i are the vertices of U in descending order of $\kappa(v^s_i)$. For this objective function the tight bound is already known to be [BGG+17]

$$\hat{f^s_{\mathrm{di}}}(U) = \max_{0<i\leq|U|} \frac{i}{|V|}\left[\frac{1}{i}\sum_{j=1}^{i} \kappa_V\left(v^s_j\right) - \bar{\kappa}_V\right]. \tag{4.8}$$

The static measure f^s_{di}, however, systematically overestimates the subgraph density, as visualised in Fig. 4.3, since it also considers the connections to the rest of the subgraph instead of just the subgraph of interest. This intuition leads to the following observation that is key for the rest of our analysis: the average coreness is monotone[1] with respect to the inducing vertex set.

[1] More formally, the average coreness is monotonously increasing over any monotonously increasing sequence of the vertex power-set 2^V, directed by the set inclusion relation $\subseteq$.

Lemma 4.1. *Let $T \subseteq U$. Then* $\bar{\kappa}_U(T) \coloneqq \dfrac{|T|}{|V|} \sum_{v \in T} \kappa_U(v) \geq \bar{\kappa}_T$.

Proof. By construction of the induced subgraphs of T and U we have $E(T) \subseteq E(U)$ and for all vertices $v \in T$ it is $N_T(v) \subseteq N_U(v)$ or equivalently $\delta_T(v) \leq \delta_U(v)$, and therefore $\kappa_T(v) \leq \kappa_U(v)$. Adding these inequalities over all vertices $u \in T$ proves the claim. □

In other words, since f_{di}^s overestimates f_{di}, we can use the tight optimistic estimator $\hat{f}_{\mathrm{di}}^s$ of f_{di}^s also as an optimistic estimator of our own measure, albeit a non-tight one. In fact, this tight bound of f_{di}^s can be derived with little work from the existing literature [BGG+17], once we use as target property the core values of G. This gives

$$\max_{T \subseteq U} f_{\mathrm{di}}(T) \leq \max_{T \in U} f_{\mathrm{di}}^s(T) = \max_{0 < i \leq |U|} \frac{i}{|V|} \left[\frac{1}{i} \sum_{j=1}^{i} \kappa_V(v_j) - \bar{\kappa}_V \right], \tag{4.9}$$

where $v_1, \ldots, v_{|V|}$ are the vertices of V ordered in decreasing core value. Notably, this bound can be adjusted to accommodate the trade-off parameter γ.

However, the bound of Eq. (4.9) only considers the core values of the entire graph, which we showed to overestimate the coreness of the induced graph in ¸Fig. 4.3; therefore, this bound can be rather loose in practice (see Section 4.5), and so we limit its use solely as a baseline. We hence derive an improved optimistic estimator for our objective function that considers the coreness of the induced subgraph, and additionally becomes tight under common conditions. At the core of this optimistic estimator lies a tight upper bound for the total coreness κ_U of Eq. (4.4) over all subsets of U, as we derive next.

More specifically, we can compute the maximum of the total coreness κ_T that can be attained over all subgraphs $G[T]$ induced by a vertex subset $T \subseteq U$. We decompose this maximum computation as

$$\kappa_U^* \coloneqq \max_{T \subseteq U} \kappa_T = \max_{1 \leq i \leq |U|} \kappa_U^i \,, \tag{4.10}$$

where we first maximise over those $T \subseteq U$ with cardinality i

$$\kappa_U^i \coloneqq \max_{T \subseteq U \,,\, |T| = i} \kappa_T(T) \;. \tag{4.11}$$

To compute the maximum κ_U^i over all subsets with fixed cardinality we first arrange all vertices $v_1, \ldots, v_{|U|}$ of U in order of decreasing coreness

$\kappa_U(v_i)$ and then we observe that κ_U^i is upper bounded by the partial sums $\hat{\kappa}_U^i = \sum_{j=1}^{i} \kappa_U(v_j)$.

We now study the sequence of these partial sums $\hat{\kappa}_U^i$ as follows. Due to their ordering, the vertices are selected one k-shell of $G[U]$ at a time in decreasing order of k, so that within each k-shell the value of $\hat{\kappa}_U^i$ increases by a constant k. This constant changes right after each k-shell (or equivalently, k-core) is exhausted. There are $K_U + 1$ such **complete core addition (CCA)** indices: each corresponds to exhausting the vertices of a k-core and thus coincides with the size of a k-core. We denote these as $n_k := |E_U(k)|$ for each k-core $0 \leq k \leq K_U + 1$.

Note that $\hat{\kappa}_U^i$ increases linearly between two consecutive complete core addition indices $n_{k+1} \leq i \leq n_i$ by exactly k. Thus, $\hat{\kappa}_U^i$ is a piece-wise linear sequence in i, whose pieces switch at indices $i = n_k$. The value of $\hat{\kappa}_U^i$ at each such index can be computed as the cumulative sum of k-shell sizes, each weighted by k; to compute the rest we use linear interpolation:

$$\hat{\kappa}_U^i := \begin{cases} \sum_{\lambda=k}^{K_U} \lambda(n_\lambda - n_{\lambda+1}) & i = n_k \\ & 0 \leq k \leq K_U \\ \dfrac{(i - n_{k+1})\hat{\kappa}_U^{n_k} + (n_k - i)\hat{\kappa}_U^{n_{k+1}}}{n_{k+1} - n_k} & n_{k+1} \leq i < n_k \\ & 0 \leq k \leq K_U \ . \end{cases} \tag{4.12}$$

Since $\hat{\kappa}_U^{n_k} = \hat{\kappa}_U^{n_{k+1}} + k(n_k - n_{k+1})$, the above is simplified as

$$\hat{\kappa}_U^i = (i - n_{k+1})k + \sum_{\lambda=k}^{K_U} \lambda(n_k - n_{k+1}) \ , \qquad n_{k+1} \leq i \leq n_k. \tag{4.13}$$

Equation (4.13) reveals $\hat{\kappa}_U^i$ to be piece-wise linear (and concave) function due to the monotonically decreasing increments k. This is demonstrated in Fig. 4.4a. Each element of the sequence $\hat{\kappa}_U^i$ can now serve as an upper bound for the maximum total coreness κ_U^i over all subsets of U with fixed cardinality i.

Proposition 4.2 (Piece-wise Linear Estimate)**.** *For the piece-wise linear function of Eq.* (4.13)

i) $\kappa_U^i \leq \hat{\kappa}_U^i$, *for all* $0 \leq i \leq |U|$
ii) $\kappa_U^i = \hat{\kappa}_U^i$, *for* $i \in \{0, n_0, \ldots, n_{K_U}\}$

— For the proof see Appendix A.2.

Using the first part of Proposition 4.2 we can upper bound the value of

f_{di}^s over all subsets of U with cardinality i by

$$\hat{\phi}_U(i;\gamma) \coloneqq \left(\frac{i}{|V|}\right)^{1-\gamma} \left(\frac{\hat{\kappa}_U^i}{i} - \bar{\kappa}_V\right)^{\gamma} . \tag{4.14}$$

Hence, the solution of Eq. (2.27) for $f_{\text{di}}(U;\gamma)$ can be written as

$$\max_{T\subseteq U} f_{\text{di}}(T;\gamma) \leq \hat{\phi}_U^*(\gamma) \coloneqq \max_{0<i\leq|U|} \hat{\phi}_U(i;\gamma) . \tag{4.15}$$

Finally, we replace Eq. (4.14) into the one above and then use Proposition 4.2 (part ii)) to show that our final bound is tight. To share intuition of its tightness, we plot the proposed optimistic estimator for $\gamma = 1/2$ (Fig. 4.4b). Note that as γ deviates from this value, the guarantee for tightness can be lost (Fig. 4.4c).

Corollary 4.3 (Optimistic Estimate)**.** *The quantity $\hat{\phi}_U^*(\gamma)$ is an optimistic estimator of $f_{di}(U;\gamma)$. In addition, $\hat{\phi}_U^*$ becomes tight for $\gamma = 1/2$.*

$$\hat{\phi}_U^*(\gamma) \coloneqq \max_{0<i\leq|U|} \left(\frac{i}{|V|}\right)^{1-\gamma} \left(\frac{\hat{\kappa}_U^i}{i} - \bar{\kappa}_V\right)^{\gamma} . \tag{4.16}$$

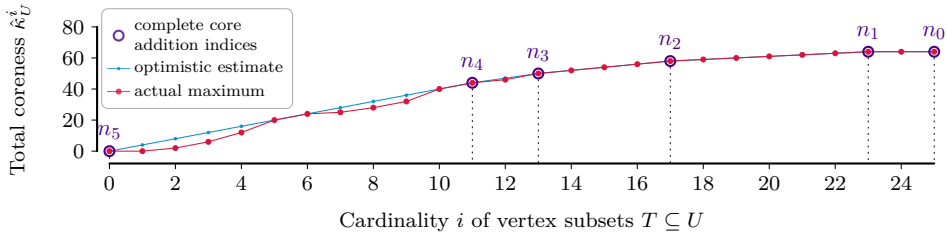
— For the proof see Appendix A.2.

Our proposed bound of Eq. (4.16) can be computed in linear time: the k-core decomposition of G is $O(n)$ [BZ03], and the maximum in Eq. (4.16) compares $|U|$ values, each computable in $O(1)$.

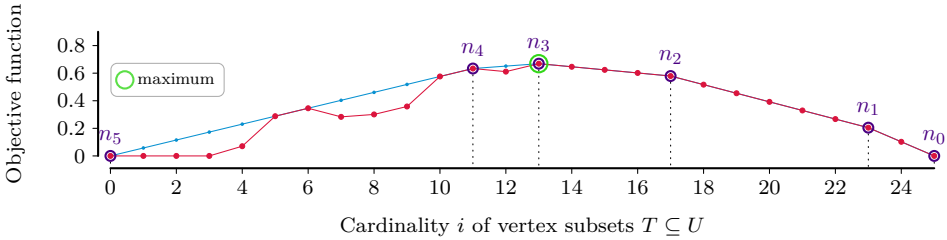
4.4 Related Work

Dense Subgraphs and Communities. The typical objective in *dense subgraph discovery* is to find the subset of vertices in a non-attributed graph that induces the subgraph with the highest edge-to-vertex ratio, a measure that has been shown to accept both an exact max-flow based polynomial time optimisation algorithm [Gol84] and a greedy 2-factor approximation [Cha00] with linear complexity. Extending this rather simplistic measure, an abundance of works reinterpret density to take into account structural information, for instance, high triangle counts [Tso14], measures based on large and/or dense k-cliques [Tso15], quasi-cliques [TBG+13], k-plexes, k-clubs, and k-cores [SEF16], just to name a few (for a survey see [LRJ+10]).

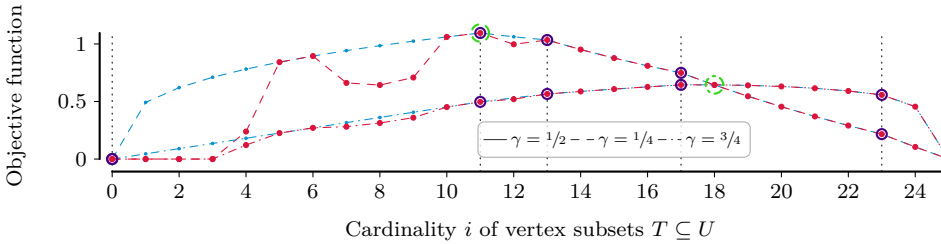
In the related yet different *community detection*, one imposes the additional constraint that the discovered subgraph must be disconnected with



(a) The maximum total coreness κ_U^i over fixed cardinality subsets is upper bounded by $\hat{\kappa}_U^i$. The latter is a piece-wise linear function of i with slope k between the consecutive CCA indices $n_{k+1} \leq i \leq n_k$.



(b) The (squared) maximum objective value over fixed cardinality subsets and its optimistic estimator $\hat{\phi}_U^*(i; 1/2)^2$. The latter is a piece-wise linear function of i; as such, its maximum over i must lie on one of the part intersections: the CCA indices. Since at these indices the bound is tight, tightness of the bound over all i follows (Corollary 4.3).



(c) As the parameter γ deviates from the value of $1/2$ the (appropriately scaled) bound loses the above piece-wise linear property. It thus becomes possible that the maximum over all i lies on a non-CCA index, where the bound is not guaranteed to match the actual maximum; then tightness may be lost. Here shown the bounds $\hat{\phi}_U^*(i; 1/4)^4$ and $\hat{\phi}_U^*(i; 3/4)^4$ with their respective maxima.

Figure 4.4: Optimistic bounds for the total coreness and the objective value over all subsets $T \subseteq U$ with fixed cardinality $|T| = i$. Here computed for the full vertex set $U = V$ of the example graph in Fig. 4.2. The bounds of both quantities coincide with the respective maximum value at each complete core addition (CCA) index $i \in \{n_5, \ldots, n_0\}$ (circled points).

Method	Task	Desc.	Model	Exact	Structure
SD-Map* [ADM16]	CD	✓	local	✓	✓
DCM [PBV14]	CD	✓	local	✓	
CoPaM [MCR+09]	CD		global		✓°
Gamer [GFB+10]	CD		global		✓°
PICS [ATM+12]	DSD		global		
AMEN [PA18]	CD		local		
LDENSE [GGT16]	DSD	✓	mixed	✓	
SCPM [SMZ12]	DSD	✓	local		✓
RoSi (ours)	DSD	✓	local	✓	✓

° The structure is only used to compare against a threshold and does not further partake in the quality of the subgroup.

Table 4.1: Comparison of related work.

the rest of the graph, which usually incurs the need for combinatorial optimisation [FH16]. Note that RoSi *solves the former* task, by adapting a k-core–based measure for mining *named vertex subsets*.

Moving on to methods which use graph attributes, we first classify them as those using graph attributes to steer a density optimisation scheme toward **cohesive subgraphs**, i.e., subgraphs with similar attributes, or others that seek the densest out of a set of subgroups, to which RoSi also belongs. *Cohesive Subgraphs.* CoPaM [MCR+09] applies *subspace clustering* on the vertex attributes to find maximal connected subgraphs that contain vertices with similar attributes, whose density surpasses a given threshold. Similarly, Gamer [GFB+10] discovers non-redundant sets of subgraphs, which must be connected γ-quasi-cliques for a given parameter γ. Note that for both methods the respective density score only needs to surpass a user-defined threshold and does not contribute to the quality of each subgraph any further. More recently, AMEN [PA18] introduces an attribute-aware variant of the established modularity measure [FH16] to detect ego-net–shaped communities with similar attributes. These last three methods score each mined pattern individually. In contrast, the *subgraph clustering* PICS of [ATM+12] uses low entropy splits of the binary adjacency and attribute matrices to form vertex clusters with similar concentration of edges and binary features. We compare RoSi to the most recent works of both subspace and subgraph clustering.

	Nodes		Edges				
Name	**No.**	**Kind**	**No.**	**Kind**	**Attr.**	α	$\bar{\kappa}$
Facebook°	4037	user	170174	friendship	20	1	52.1
Google+°	78393	user	28312689	friends	10	0.1	366.7
Delicious•	1867	user	15328	contact	50	0.3	11.0
Lastfm-Artists•	1892	user	25434	artist	15	1	14.6
Twitter°	51246	user	1735925	follower	14	1	35.7
DBLP	17488	author	97070	co-auth.	113	0.3	8.5
IMDB	23700	crew	1134676	collab.	55	0.8	50.9
GATTWTO	177	country	230777	trading	27	1	1606.7°
Amazon°	16641	record	162815	purchase	145	0.7	13.9
Lastfm-Songs•	251272	song	1179317	similarity	50	0.5	5.2

Sources: °SNAP repository •HETREC Workshop •Million Song Dataset
°Multi-graphs may have degeneracy $K \geq |E|$.

Table 4.2: Overview of dataset statistics.

Subgroup Discovery. As a subgroup discovery task, RoSi belongs to the data driven methods that use an implicit target concept. That means that the measure of exceptionality is not based on a target variable that is explicitly listed in the dataset, but is instead computed implicitly based on the robust connectedness property of the subgroup entities. Perhaps the closest to our work within this family is SCPM [SMZ12] which uses a structured density measure based on quasi-cliques that must be sampled from each subgraph to estimate how many of its vertices these cliques cover. This method needs many hard-to-specify parameters, is only approximate and, as our experiments show, slower than RoSi. Although faster, LDENSE [GGT16] is a *greedy* search for the describable subgraph with the highest typical density (i.e., edge-to-vertex ratio). Less related are methods solving the community detection problem, instead. For instance, Atzmueller et al. [ADM16] also use a branch-and-bound scheme for exhaustive search using target concepts for community detection, such as a local variant of modularity that is computed only on the subgraph (LMDL) and the inverse conductance (COIN).

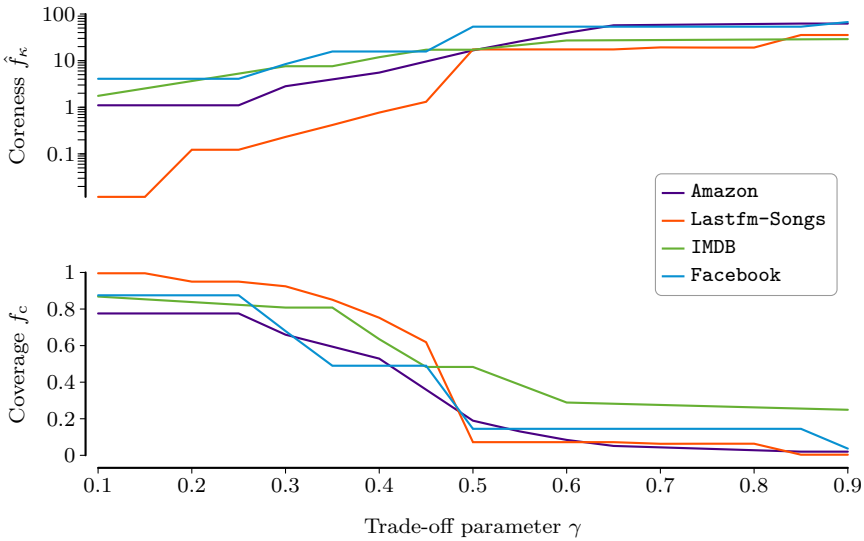


Figure 4.5 [Coreness vs. Coverage]: Increasing the coreness–coverage trade-off parameter γ yields smaller but more robustly connected subgraphs.

4.5 Experiments

In this section we empirically evaluate RoSi.[2] We consider 10 datasets that together span multiple domains and different kinds of represented entities and relations from public sources with up to thousands of vertices and millions of edges. These consist of both graphs and multi-graphs, and describe various types of networks: social, similarity, co-occurrence, collaboration networks, among others.

For the needs of our comparisons we implemented and embedded in our own framework the methods of LDENSE, COIN and LMDL. We also used the provided sources for the methods PAICAN, SCPM, PICS and AMEN. All experiments were run for at most 36 hours, similarly to the time allowed for our own algorithms. Exceptions include the PAICAN algorithm which failed to converge[3] in all of our datasets, possibly due to their size.

Where more parameters were required, we experimented with several settings and chose the best results overall. Such a case was not only COIN, for which we tried several minimum supports, SCPM, where we experimented with clique sizes of 4, 10 and 100 while kept minimum supports as the ones

[2]Code and data are available at https://eda.mmci.uni-saarland.de/prj/rosi/.

[3]More specifically, after either the first, or just a few iterations, the yielded ELBOW of the score degenerated to a value of `nan`.

reported in the presenting paper [SMZ12]. One problem of this method was the overwhelming number of results that it was providing (> 100), or the complete lack thereof, and we seemed to be lacking intuition on what good values would be for a dataset a priori. In the section below, we report the results of this experimental procedure.

4.5.1 The Generality–Connectedness Trade-Off

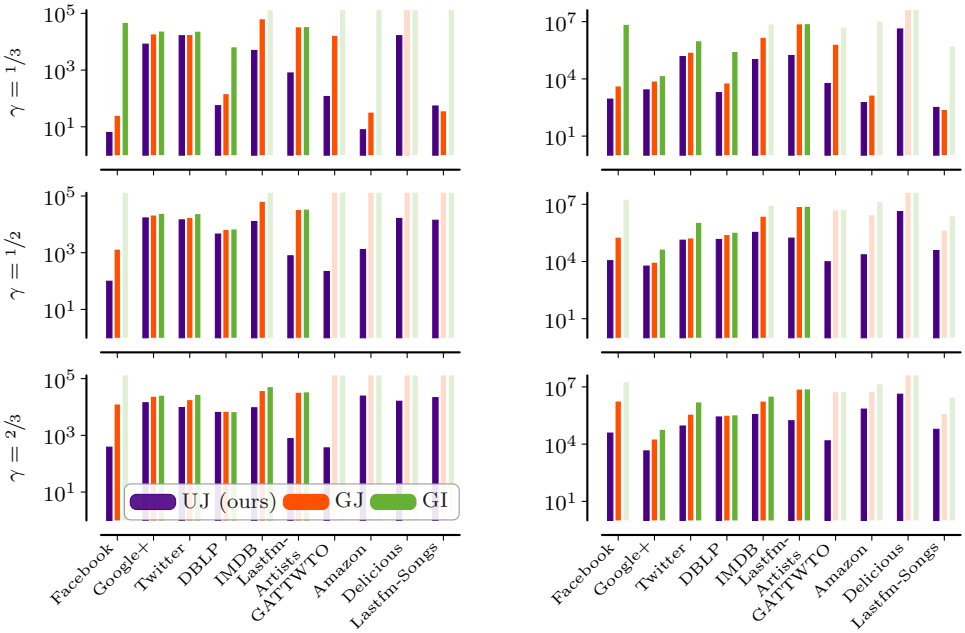
We next demonstrate the effect of the trade-off parameter γ, which offers at once a smooth and intuitive mechanism to tune the importance between the size (coverage) and the connectedness (density) of the discovered subgraph. We study datasets with highly diverse base predicates that allow the greatest flexibility in the resulting descriptions, and mine the top result for increasing values $\gamma \in \{0.1, 0.15, \ldots, 0.9\}$ and plot the coverage and connectedness of the topmost result (Fig. 4.5).

Continuously increasing parameter γ leads to smaller and more densely connected subgraphs—it thus intuitively steers the results toward more general or more connected subgraphs, and allows us to explore a compact subset of the Pareto frontier defined of the two scores of generality and density, as discussed in Section 2.5. Thus, the output of this method is a collection of subgroups for each dataset, as shown in Table 4.3.

4.5.2 Pruning Efficiency

We first study how the efficiency of RoSi is affected by the pruning potential of the proposed *induced-joint* (4.16) (UJ) against the *global-joint* (4.9) (GJ), which is the standard algorithm available from literature and here serves as a baseline. For the experiments we use the default trade-off parameter of $\gamma = 1/2$, along with the representative choices of $\gamma = 1/3$ and $\gamma = 2/3$, corresponding to increasing the importance of coverage and robust connectedness, respectively. When the computation time (for UJ) exceeds 7 hours, we lower the approximation factor α by 0.1 or decrease the depth limit, favouring a deeper search when possible. We report the wall-clock times and traversed nodes in Fig. 4.6.

We first focus on the running times (Fig. 4.6a) to note that with one exception $\hat{f}_{\mathrm{UJ}}$ outperforms the baseline, which often does not even terminate. This becomes even more pronounced once for any given dataset we consider the performance across all 3 selected values of the trade-off parameter: in this case there are several datasets for which $\hat{f}_{\mathrm{UJ}}$ achieves at least an order of magnitude better performance for some value of γ.



(a) Wall-clock running time (s). **(b)** Visited nodes during search.

Figure 4.6 [Lower is Better]: Efficiency of the optimistic estimators: higher pruning efficiency translates to less expanded nodes and thus shorter running times. Experiments exceeding a runtime of 36 hours (dotted line) are faded out.

This higher efficiency of the proposed bound is due to its higher pruning capacity. Indeed, both optimistic estimators have linear time complexity, and thus the only factor that differentiates the performance of the two is the number of expanded nodes during search (Fig. 4.6b). Since for each dataset the order of predicates P is fixed, without pruning, the sequence of traversed nodes would be the same. Within this sequence, pruning allows to skip sub-optimal branches of nodes, which leads to smaller or larger advances, when the bound is looser or tighter, respectively. Importantly, during approximate search ($\alpha \ll 1$) pruning becomes overzealous: then a bound might skip a good node, which a looser bound would "fail" to skip; this occasionally leads to an advantage for the looser bound, later on. This is more likely to occur as α lowers. In our experiments this only happens for `Lastfm-Songs` ($\gamma = 2/3$) when the baseline gains a slight advantage, which is no surprise given the low approximation factor of 50%.

As a side-note, we also observe that a deeper search depth allows a better margin for a tighter optimistic estimator to thrive. Indeed, the only case

where the pruning superiority of $\hat{f}_{\mathrm{UJ}}$ is not pronounced is for the DBLP dataset (for $\gamma = 2/3$), where the algorithm only reaches a depth of 3, the lowest among all datasets (Table B.1).

Overall, the experiments corroborate the theoretically expected superior pruning of the proposed $\hat{f}_{\mathrm{UJ}}$ and show that it readily translated to shorter running-times. Note that RoSi is a variant of IDDFS, and as such is very frugal in its memory usage. If used in a BFS setting, a looser bound would run into memory issues much faster than a tighter one.

4.5.3 Connectedness versus Density

Here we compare RoSi to representative works in terms of both our proposed robust connectedness and also typical density (edge-to-vertex ratio). Although our task is dense subgraph discovery, we also compare against more loosely related approaches for community detection.

We first compare against state-of-the-art methods which describe the found patterns: LDENSE [GGT16], SCPM [SMZ12], and two target concepts for community detection from subgroup discovery on graphs: COIN [AM11]) and local modularity (LMDL [ADM16]). We plot the best results of each method in Fig. 4.7a. RoSi scores the highest in terms of robust connectedness, while in terms of density it is on par with the rest.

We further compare RoSi with two recent methods for cohesive subgroups: PICS and AMEN, neither of which do not provide descriptions. Since these methods output several patterns, we show all discovered vertex sets in the Pareto front of the two metrics (Fig. 4.7b) with empty circles, designating the absence of a description. Although rarely, other methods may score a higher density and even robustness than RoSi, as their optimisation not constrained. To put them in perspective, however, we further mine the closest subgroup in terms of Jaccard distance to the one provided by each algorithm, and link to it the unconstrained solution with an arrow. As expected, these solutions score lower than those of RoSi.

4.5.4 Intelligible Subgraph Descriptions

To qualitatively assess our results we mine the top describable subgraph for the IMDB dataset which offers attributes that are easily interpretable, consisting of movie cast members which are connected when they have collaborated in at least one movie. In Table 4.3a we track the top description over varying steps of $0.1 \leq \gamma \leq 0.9$.

Starting with larger subgraphs (high γ) we read the first result as: *the drama movie cast has a robust connectedness of* 1.8 *collaborations on average*

more than what is usual in the entire industry. Moving into denser graphs, we find that established actors (i.e., debuting before '96) collaborate more with each other. Here, also a negated predicate is informative: the London BFI festival is known to nominate more diverse films, with cast harder to have collaborated with each other, therefore removing it increases connectedness. We further find that additionally producing a movie in the US leads to substantially higher connectedness. Overall, the discovered patterns reveal an interpretable story.

Similarly, Table 4.3b lists discovered subgroups from the Lastfm song similarity dataset. These reveal that the few live recordings are dissimilar, most likely due to the higher noise levels involved in live venues. We also find out certain genres to offer greater variation in their songs, e.g., metal, indie, experimental, punk, and alternative rock, exemplary genres known for novel sounds and breaking norms. Lastly, we identify genres with many more similar pieces within them than the average, for instance the 18 thousand oldies and the thousand dance-party 70's songs.

We also report selected informative subgraphs discovered from another 4 datasets (Table 4.3c). Interesting findings include that the `Google+` social network contains a community of photographers, which have 140 other photographers as friends on average more than the dataset average; similarly, in Twitter, the followers of the American artist Hayley Williams are exceeded by 120 connections the average connection in the dataset, and even more so fans who refer to YouTube itself. From the `DBLP` dataset we notice that the people publishing in the ICDM conference have a slightly higher tendency to cite other people of the same field. Finally, finally the discoveries of the `GATTWTO` dataset show that the member countries of the GSP trade agreement use on average 253 more trade routes ore than what is usual.

4.6 Discussion

Our experiments show RoSi to be feasible even on large graphs and to yield meaningful and easily interpretable results. Nevertheless, it comes with weak points, discussed below.

Our measure leverages the structural properties of k-cores to create a coherence measure suitable for subgroup discovery. However, an easily shown property of a k-core $H(k)$ is that if its vertex count n is relatively higher than its core number $n \geq \lambda(k+1)$ then it can consist of up to λ k-core components $H_k^1, \ldots, H_k^\lambda$, all of which are disconnected between them. Therefore, our measure does not enforce that the described entities are *all* (well) connected together. However, this is no shortcoming of our method, as our goal is not

to name individual connected components or communities. Instead we seek to discover factors which lead to higher coherence of—possibly separate—groups of entities, in terms of conjunctions of vertex attributes, which is well achieved.

This caveat means that, for instance, our discovery from the `Google+` network (in Table 4.3c) does not indicate that all of the 2835 photographers are friends with each other, which would be a very hard constraint to satisfy. Rather, this discovery means that photographers tend to form robustly connected clusters, whose average minimum edge-connectedness exceeds the network average by 140. Should connected subgraphs be required, this can always be enforced as a post-processing step.

Importantly, despite relying on first order moments of core values, our measure is not particularly susceptible to outliers. This is due to several graph dynamics that together counter-act the outlier sensitivity. First, higher k-cores need more vertices to "collaborate" together, and thus higher k-cores purely due to outliers become increasingly unlikely. Second, low k-cores are unlikely to occur in the results due to outliers because our score aims to find subgraphs with deviating coreness, and low-coreness graphs are highly likely due to the power-law-like distribution of coreness in real-world graphs, which comes in addition to coreness being positive.

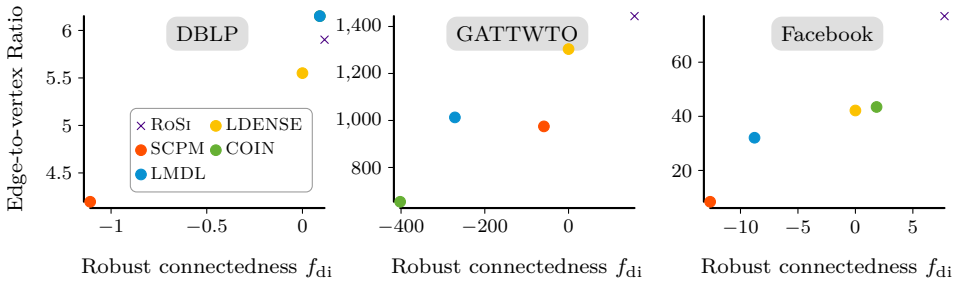
4.7 Conclusion

In this chapter we studied the problem of finding robustly connected subgraphs that are easily described. For this we first formally define the concept of robust connectedness as the property possessed by graphs that contain node clusters that are difficult to shatter; we then propose a measure for this property based on the k-core decomposition of a graph.

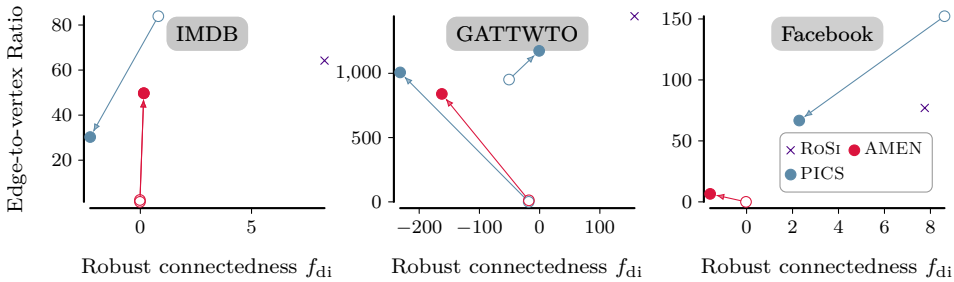
We subsequently adapt our novel measure for its use in subgroup discovery, which gives rise to the coreness impact function, that incorporates the notions introduced in Section 2.3. We additionally provide an efficient algorithm to compute its tight optimistic estimator, which enables the use of our IDDFS algorithm to efficiently optimise our objective.

Our experiments show that, although our problem is inherently exponential, RoSi can analyse real-world graphs with up to millions of edges and tens of thousands of vertices within reasonable time. This is largely to the efficiency of the tight optimistic estimator that we derived for our measure, which is only of linear computational complexity. This makes our optimistic estimator optimal, in the sense that it does not change the complexity class of each iteration, as discussed in Section 2.4.2.

Importantly, our results are meaningful and come equipped with intelligible descriptions, which establish its clear superiority when compared with alternative methods. These alternative methods either provide no descriptions, or only provide approximate or implicit ones.



(a) Comparison to methods which provide descriptions.
Among all subgroups, RoSi is always the rightmost, as it finds the optimum one in terms of robust connectedness.



(b) Description-less methods: a hollow mark designates their result: a subset of entities without a description. Arrows point to the closest (describable) subgroup in terms of Jaccard similarity.
The (unnamed) subsets that these methods find can score higher than our result in both density notions, as in case of the `Facebook` dataset [right], since they are unconstrained. However, post-hoc fitting of the closest description gives an unpredictably scoring subgroup, which is always worse than that of RoSi in terms of robust connectedness, since our method computes the optimal subgroup in this regard. In our experiments, these methods happen to score lower than our result also in terms of vertex-to-node ratio.

Figure 4.7 [Upper Right is Better]: Comparison of the two metrics for density: edge-to-vertex ratio and our robust connectedness, on different datasets. RoSi is not only able to discover subgroups with the highest robust connectedness, as expected, but it also scores on par with competing methods also in terms of the typical edge-to-vertex ratio.

γ	Drama	Comedy	¬BFI	Debut $\leq$ '96	Debut $\leq$ '05	US	Movies	Dens.
$[0.10 - 0.30]$	✓						20,579 (86.8%)	1.8
$[0.30 - 0.40]$			✓		✓		19,150 (80.8%)	7.6
$[0.40 - 0.45]$			✓	✓	✓		15,057 (63.5%)	11.9
$[0.45 - 0.60]$	✓	✓	✓	✓	✓		11,455 (48.3%)	17.1
$[0.60 - 0.90]$	✓	✓	✓	✓	✓	✓	6,843 (28.9%)	27.1

(a) Discovered subgraphs from `IMDB`.

γ	Description	Songs	Dens.
$[0.10 - 0.20]$	[¬seen_live]	250,168 (99.6%)	0
$[0.20 - 0.30]$	[¬experimental]	238,682 (95.0%)	0.1
$[0.30 - 0.35]$	[¬metal]	232,272 (92.4%)	0.2
$[0.35 - 0.40]$	[¬metal]$\wedge^{\ddagger}$	213,803 (85.1%)	0.4
$[0.40 - 0.45]$	[¬experimental]$\wedge$[¬metal]$\wedge^{*}$	189,054 (75.2%)	0.8
$[0.45 - 0.50]$	$^{\circ}\wedge$[¬ambient]$\wedge^{\bullet}\wedge$[¬metal]$\wedge^{\ddagger}\wedge$ [¬alternative_rock]$\wedge$[¬punk]$\wedge^{*}$	155,446 (61.9%)	1.3
$[0.50 - 0.70]$	[oldies]	18,089 (7.2%)	17.4
$[0.70 - 0.75]$	$^{\circ}\wedge$[¬90s]$\wedge$[oldies]$\wedge$[¬hard_rock]	15,842 (6.3%)	19.2
$[0.75 - 0.85]$	[¬00s]$\wedge$[oldies]$\wedge^{\bullet}\wedge^{*}$	15,953 (6.3%)	19.1
$[0.85 - 0.90]$	[¬seen_live]$\wedge$[party]$\wedge$[70s]$\wedge$[dance]	862 (0.3%)	35.5

$^{\circ}$[¬seen_live] $\wedge$ [¬experimental] $^{\bullet}$[¬hard_rock] *[¬indie] ‡[¬indie_rock]

(b) Discovered subgraphs from dataset `Lastfm-Artists`.

Dataset	γ	Description	Entities	Dens.
Google+	$[0.10 - 0.90]$	[job=photographer]	2,835 (3.6%)	138.9
Twitter	$[0.10 - 0.85]$	[@yelyahwilliams!]	740 (1.4%)	119.9
DBLP	$[0.10 - 0.35]$	[vna:ICDM]	9,022 (51.6%)	0.1
GATTWTO	$[0.25 - 0.55]$	[tra:GSP]	110 (62.1%)	253.5

(c) Discovered subgraphs of special interest.

Table 4.3: Discovered subgraphs for different trade-off parameters γ.

5 Kernelised Subgroup Discovery

In this chapter we revisit the core goal of (data-centred) subgroup discovery: finding the describable sub-population of entities whose distribution is the most deviating from what is considered usual. As we saw, the existing methods in this field have only quantified this deviation for simple target variables that were almost exclusively scalar. Thus, the use of the existing methods is limited to datasets with little to no structure. Instead, in this chapter we will treat entities that are each associated with an arbitrary structure. This relaxation allows for a much more flexible setting than what just a scalar attribute could accommodate. In fact, within the previously studied framework, we adopt a new perspective in which the structure associated with each entity becomes, itself, the target variable.

Equivalently, we consider datasets whose entities themselves are arbitrary structures; these structures can be represented in a flexible form and consist of multiple dimensions, that can even vary from one entity to the other. Such entity structure can be, for example, proteins, molecules, graphs, images, time series, or effectively any out of the limitless variety of structures on which a meaningful positive definite kernel can be defined. As usual, we also assume that a set of relevant attributes is available for each of the given entities. More specifically, these attributes should (indirectly or explicitly) capture interesting traits of the studied entities, using which it is possible to meaningfully group them. Our aim is, hence, to extend the applicability of typical subgroup discovery on such datasets. Applying such a method we can then find that particular named subset whose structure deviates as much as possible, either from the rest of the dataset or from the entirety of it.

As an example, consider the active research field of computer-aided drug discovery, where molecules are scrutinised based on their structure and general chemical properties, in order to find potential drug candidates. A great amount of information on chemical properties for molecules is available, or can be inferred with relative ease via simulations based on the molecule shape, e.g., electronic density, number of atoms, benzene rings, etc; importantly, such chemical and structural properties of molecules can be captured by an appropriate kernel on molecule shapes. However, only a relatively small set of drug-like substances has been meticulously annotated with their drug-related properties by lab specialists. These latter properties could be suggestive

(a) Entities of a toy dataset with structure a point in $\mathbb{R}^2$ [left]. These points lie on a low-dimensional manifold, as revealed when we project them along the first 2 eigenvectors of a Gaussian kernel Gramian [right].
Alongside are shown the validities of two predicates p_1 and p_2.

(b) The subgroup $p_1 \wedge \neg p_2$: the most exceptional within the subgroup language of p_1 and p_2. We measure the exceptionality of a subset in Hilbert space, as the difference of the subset mean (coloured mark) to the dataset mean (black star). Although the marked cluster is more exceptional, it has no description and is not considered.

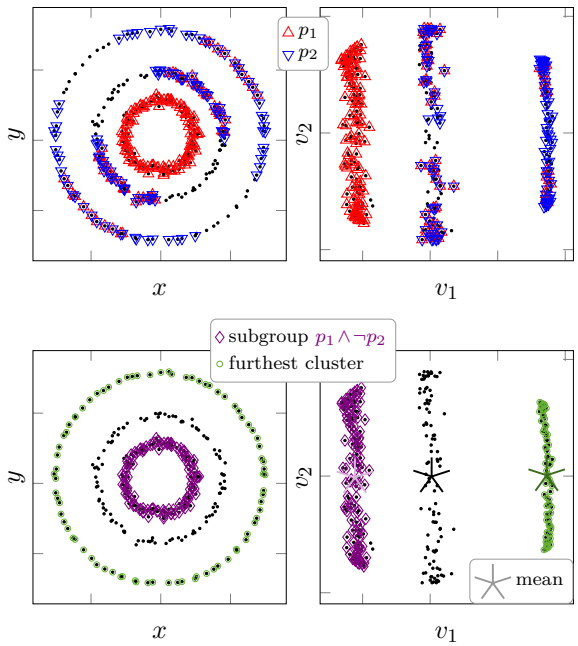


Figure 5.1: Toy example of points with structure in $\mathbb{R}^2$ along with their two predicates p_1 and p_2, that model suggestive traits (a). Although the human eye can easily distinguish clusters in this toy dataset, this is not trivial for a machine learning model. We therefore measure the exceptionality in Hilbert space, where its value can be easily measured as the difference of subset means (b).

traits like toxicity, bio-availability, affinity to a specific target, etc, which can be highly indicative of the fitness of a substance for a medical condition. It would then be of great use to find groups within the available drug dataset that stand out with respect to their shape-based properties, but that at the same time share a common set of drug-related, revealing traits. Thus, by harnessing the power of subgroup discovery, in this paradigm we shift from the discovery of a list of molecules, like paracetamol, ibuprofen, etc, to, instead, a set of common traits of them that intelligibly describe this set; for instance, such a finding could be "*'painkillers with high bio-availability and low toxicity describe drugs with outstanding molecular properties'*". The usefulness of such a discovery can be that important deviations within a known class of drugs can be a target of further scrutiny, that focuses on conspicuously deviating parts fo this class, or give indications of potential improvement of pharmaceutical properties.

This paradigm deviates from having one single studied suggestive trait

like toxicity, for which one could typically fit a classifier on a set of structure-related attributes, as is achievable with several methods from supervised learning. Instead, in our setting we search for these subsets of molecules which are interesting due to a difference in general structure, as assessed by a positive definite kernel, while always retaining an intelligible description. More specifically, we generalise subgroup discovery to find the subgroup whose average representation in the associated Hilbert space is most deviating from the representative of the usual data. Consider, for example, the toy dataset of entities with structure in a two-dimensional vector in $\mathbb{R}^2$, which is depicted in the top-left of Fig. 5.1a. By a quick glance on the arrangement of these points, one can probably already guess that these points have a low-dimensional structure, and in fact one that can be very well captured by a Gaussian kernel; indeed, this structure is revealed once we project these points along the first two eigenvector directions of the kernel Gramian. We also assume the availability of two suggestive traits, which we mark with the predicates p_1 and p_2, respectively. In this scenario, although finding exceptional subgroups is very problematic in the original space, using the Gaussian kernel we can perform the same task quite easily in the Hilbert space. Doing so, we find the subgroup $p_1 \wedge \neg p_2$ (depicted in Fig. 5.1a) which is the optimal among all those subsets that have a description.

It also becomes important to define what are the structural properties that the kernel should deem relevant for its similarity assessment, or, in other words, how to choose an appropriate kernel. Typically, the kernel—or its parameters—are chosen based on methods like cross validation, which, in turn, requires the availability of a clear cut metric that should be optimised. Despite, however, the several attributes available in the dataset, there is no sensible way to single out one of them as a regression or classification variable, and therefore we cannot directly use standard metrics like regression loss or accuracy. We hence make two key assumptions:

i) The entity properties included in the dataset are axiomatically relevant to the specific application, by the mere fact that they were included to begin with; by extension, relevant are also the predicates derived from these properties.
ii) Additionally, the similarity of the entities in two subgroups, defined each by a logical expression of these predicates, is related to the similarity of these logical expressions themselves.

We use these assumptions to propose a fitness measure of a kernel for our task, that at the same time takes into consideration all available predicates derived from the entity properties. We then use this measure within established hyper-parameter optimisation methods to select a good kernel for our task,

and additionally study schemes for multiple kernel learning where simpler kernels are linearly combined into a more fit one.

Thus, the main contributions of this paper can be summarised as follows.

- We propose a family of objective functions on subsets of entities, such that a set with outstanding structural characteristics stands out, where a positive definite kernel is consulted for the similarity of these structural similarities.
- We provide an upper bound for our objective functions that can be used for its efficient combinatorial optimisation within our IDDFS algorithm.
- We provide methods to tune the hyper-parameter of the involved kernel, which takes into account all available properties of the dataset equally.

5.1 Preliminaries

In this chapter we study datasets whose entities are each associated with additional information that can be used to evaluate a similarity between these entities. This additional information replaces—and in fact generalises—the role of the target variable of typical subgroup discovery, as it can have virtually any structure, such as graphs, time series, or images, and naturally also scalars as a trivial special case.

More precisely, now the entities are assumed to have a given structure that we use to compare them through special functions $\kappa : E \times E \to \mathbb{R}$ that map any pair of entities to a real value that measures the similarity between them. These functions are called **positive definite kernels** and have useful properties, the relevant of which we briefly study below.

5.1.1 Positive Definite Kernels

Positive definite kernels are a broad family of well-behaving functions that generalise the inner product between two entities. These entities belong to the same domain $\mathcal{X}$, which can contain anything from scalars and vectors, to virtually any of much more complex objects, such as time-series, images, graphs, or molecules. In this way, the positive definite kernels define useful similarities between practically arbitrarily structured objects, which can be seen as a generalisation of the cosine similarity.

Formally, a **positive definite kernel** $\kappa : \mathcal{X} \to \mathbb{R}$ is a symmetric function[1]

[1]The field of reals is sufficient for the needs of this work. However, all these definitions carry on to the field of complex numbers $\mathbb{C}$ after simple adaptations, mainly the replacement of the transpose operation with that of the Hermitian adjoint.

that generalises positive definite functions over $\mathcal{X}$, so that for any finite $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ it must be

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \kappa(x_i, x_j) \overset{!}{\geq} 0\,, \qquad \text{for any } c_1, \ldots, c_n \in \mathbb{R} \qquad \Longleftrightarrow \tag{5.1}$$

$$\mathbf{K} \overset{!}{\succeq} 0\,, \qquad [\mathbf{K}]_{i,j} \coloneqq \kappa(x_i, x_j)\,; \tag{5.2}$$

in the more compact definition of Eq. (5.2) we used the **Gramian** matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ of X, defined as the symmetric matrix whose elements are the values of the kernel for each combination of pairs from X.

As a direct result of this definition, it can be shown that for each positive definite kernel there is i) an associated **Hilbert space** $\mathcal{H}$ of real functions and ii) a feature map $\phi : \mathcal{X} \to \mathcal{H}$, such that

$$\kappa(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}\,, \tag{5.3}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product of the Hilbert space. Equivalently, every feature map to a space of real functions defines a positive definite kernel. This latter fact has the important implication that any function that gives a representation of an object as a vector defines a positive definite kernel. This includes a multitude of examples of embeddings for text, images and other structures, and, importantly, also all representations that can be learned through a neural network pipeline.

For any finite subset $X \subseteq \mathcal{X}$, we can also represent the value of the kernel as a finite decomposition of it into the Gramian eigenvectors[2]. This essentially provides a convenient representation of the elements of $\mathcal{X}$ in the associated Hilbert space $\mathcal{H}$, that reveals their relation in a Euclidean-like distance. The most accurate low-dimensional representation of these points is provided by using the coefficients of the eigenvectors with the largest eigenvalues. We will extensively use this representation to elucidate the structural relation of the entities in the discovered subgroups of this chapter.

5.2 Most Outstanding Named Entity Subset

We now work toward shaping the intuition of our goal into a formal problem. Our task can be interpreted as the need to find the particular subset $Q \subseteq E$ that has the maximally deviating distribution from either that of the whole

[2] In fact, by Mercer's theorem there is also an (uncountably) infinite extension to this: when the domain $\mathcal{X}$ admits a well-behaving (Radon) positive measure, it also admits a spectral decomposition into no more than countably infinite eigen-functions [SS16].

dataset or from the distribution of the complement $\bar{Q}$. The first requirement is compatible with the assumption that our dataset comprises the entire population, and therefore its statistics correspond to the true statistics from which the sought after subset is to deviate. On the contrary, if we assume E to be just a sample of the true distribution of the population, then in our assessment we may not include the sample E, and must consider only the distributional distance between Q and $\bar{Q} \coloneqq E \setminus Q$. We hence refer to the first problem as **anomalous** discovery and to the latter as **contrastive**. The next step is to specify a way to measure the distributional distance between the two sets, with respect to a positive definite kernel.

5.2.1 Maximum Mean Discrepancy

Exactly for this task Gretton et al. [GBR+07] provide a special case for a known result from real analysis [Dud02, Lemma 9.3.2]. This informally states that for any two (Borel) probability measures p, q, defined over a space $\mathcal{X}$ with a metric, it is $p = q$ if and only if when we transform all elements of $\mathcal{X}$, their mean under p is equal to the mean under q, for all transformations that come from a dense enough space, like the set of bounded continuous functions over $\mathcal{X}$.

Gretton et al. [GBR+07] propose to use the unit ball in the Hilbert space of a reproducing kernel which serves as a dense enough space which satisfies the above lemma. The resulting measure between two distributions p, q is the Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD}(p, q) \coloneqq \|\mu(p) - \mu(q)\|_{\mathcal{H}}\,, \tag{5.4}$$

$$\mu(\cdot) \coloneqq \mathbb{E}_{x\sim\cdot}[\phi(x)]\,, \qquad \hat{\mu}(P) = \frac{1}{|P|}\sum_{x\in P}\phi(x) \tag{5.5}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is the feature map of a given kernel on $\mathcal{X}$ and μ is the mean of the points in $\mathcal{H}$ that the elements of $\mathcal{X}$ are mapped to through ϕ. The mean is either of the entire space under p or (approximately) over a finite subset $P \subset \mathcal{X}$ that was itself sampled under p. More relevant for our needs is the (squared) biased empirical estimator of MMD, which for two

sets Q, Q' is

$$\widehat{\mathrm{MMD}}^2(Q,Q') = \frac{1}{|Q|^2}\sum_{\epsilon,\epsilon'\in Q}\kappa(\epsilon,\epsilon') - \frac{2}{|Q||Q'|}\sum_{\substack{\epsilon\in Q,\\ \epsilon'\in Q'}}\kappa(\epsilon,\epsilon') + \frac{1}{|Q'|^2}\sum_{\epsilon,\epsilon'\in Q'}\kappa(\epsilon,\epsilon')\,, \quad (5.6)$$

where $\kappa(\epsilon,\epsilon') = \langle\phi(\epsilon),\phi(\epsilon')\rangle_{\mathcal{H}}$. We hence adopt the measure of Eq. (5.6) for the quantification of the dissimilarity of Q and its pair.

5.2.2 An Objective for our Task

In contrast, however, to the setting for which the MMD was developed, we need to evaluate as candidate sets $Q \in \mathcal{L}$ all those that result from a named combination of the dataset attributes. This means that without proper scaling, selecting just an outlier could trigger a false discovery. We therefore adapt the $\widehat{\mathrm{MMD}}$ by multiplying it with a scaling factor $a(|Q|)$, which depends only on the size of Q, and can be interpreted as a size prior. This yields our objective function

$$J(Q;\kappa,\gamma) \coloneqq a_t^{\gamma}(|Q|)\cdot\widehat{\mathrm{MMD}}_{\kappa}^2(Q,Q'_t)\,, \quad \gamma > 0\,, \quad (5.7)$$

where $t = \mathrm{ano}$ and $t = \mathrm{con}$ indicate the anomalous and contrastive assumptions, respectively, for which we define

$$\begin{aligned} a_{\mathrm{ano}}(m) &\coloneqq m & Q'_{\mathrm{ano}} &\coloneqq E \\ a_{\mathrm{con}}(m) &\coloneqq \frac{m(n-m)}{n} & Q'_{\mathrm{con}} &\coloneqq E \setminus Q\,. \end{aligned} \quad (5.8)$$

The scalar γ is a tuning parameter that controls the relative importance between the prior on the subset cardinality and the deviation component of the objective.

Note that we are not limited to using the given priors and any reasonable choice will do. Our own is based on the intuition that larger sets are less prone to be outliers and are generally more informative. In the contrastive case, due to the symmetry $J[\mathrm{ano}](Q) = J[\mathrm{ano}](\bar{Q})$ we wish for $|Q|$ to be far from both extremes, while for $t = \mathrm{ano}$, larger sets are harder to compare as structurally different from the whole dataset, and our simpler choice is enough. Additionally, in this case (and for γ=1) we can write $\sqrt{J(Q)} = \sqrt{|Q|}\|\hat{\mu}(Q) - \mu(E)\|_{\mathcal{H}}$, where $\hat{\mu}$ is as defined in Eq. (5.4). Since we also

assumed that E is the full population, $\mu(E)$ is the true mean and therefore our score resembles (the square of) the z-score of the empirical mean estimator $\hat{\mu}(Q)$, where the difference to the mean is replaced with the norm in the Hilbert space.

We now reformulate our objective to reveal its structure, make it more convenient for what follows, and to further show that both problems differ only in the choice of the set cardinality prior $a_t(|Q|)$.

Lemma 5.1. *Let $m_Q \coloneqq |Q|$ be the cardinality of any entity subset. Then we can write our objective of Eq.* (5.7) *as*

$$J(Q;\kappa,\gamma) = a_t^{\gamma-2}(m_Q)\mathbf{z}_Q{}^\top K \mathbf{z}_Q\,, \tag{5.9}$$

where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the Gramian $\mathbf{K}_{i,j} \coloneqq \kappa(\epsilon_i,\epsilon_j)$ and

$$\mathbf{z}_Q \coloneqq \mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}\,, \tag{5.10}$$

for $\mathbf{e} \coloneqq (1,\ldots,1) \in \mathbb{R}^n$ the vector of all ones and $\mathbf{x}_Q \coloneqq (\mathbb{1}[\epsilon_i \in Q])_{i=1}^n$ the characteristic vector of set Q; here we denote $\mathbb{1}[\cdot]$ the characteristic[3] function.

We can now formalise our problem as follows.

Problem 3. *Given dataset E with attributes yielding the predicates P and with structure captured by kernel κ, solve*

$$\max_{Q\in\mathcal{L}} J(Q;\kappa,\gamma)\,. \tag{5.11}$$

This is a hard combinatorial problem and can be solved optimally by the classical Branch and Bound algorithm which we list in the supplementary material. For the efficient use of this algorithm, however, an appropriate upper bound is necessary.

5.2.3 An Upper Bound for our Objective

We now derive an upper bound for Eq. (5.9) that can be computed in linear time, assuming an one-time sorting operation with a time complexity of $O(n \log n)$.

Formally, we seek a function $\hat{f} : 2^E \to \mathbb{R}$ that when evaluated at an entity subset $Q \subseteq E$ computes an upper bound of the objective over all subsets of its argument, $\hat{f}(Q) \geq \max_{R\subseteq Q} f(R)$. Such a bound can be computed in two steps: first we can bound the objective exclusively over all subsets $R \subseteq Q$

[3] That is, $\mathbb{1}[\cdot] = 1$ if the condition $\cdot$ is satisfied and 0 otherwise.

with a fixed cardinality m_R, and then we can compute an upper bound for all subsets as the maximum of all cardinality-constrained maxima. We hence seek an upper bound of the sub-problem

$$\hat{f}_t(Q;\kappa,\gamma,m) \geq \max_{R\subseteq Q\,,\,|R|=m} J(R;\kappa,\gamma)\;. \tag{5.12}$$

Since now the size m_Q remains constant, we can derive a bound for each sub-problem as follows. Let $\mathbf{e}_i$ denote the i-th vector of the standard basis, i.e., the vector with a single one at the i-th position, and define $\mathbf{e}_{:m} \coloneqq \sum_{i=1}^{m} \mathbf{e}_i$. Let $\mathbf{v}_1,\ldots,\mathbf{v}_k$ be the eigenvectors of $\mathbf{K}$ with rank $k \leq n$ and corresponding eigenvalues $\lambda_1 \geq \ldots,\lambda_k$. Further, denote $\mathbf{v}_{i\uparrow[Q]}$, $\mathbf{v}_{i\downarrow[Q]}$ the vector with those entries of $\mathbf{v}_i$ for which the characteristic function of $\mathbf{x}_Q$ is non-zero, sorted in increasing and decreasing order, respectively.

Lemma 5.2. *Given any integer constant $\rho < k$, an upper bound for the problem in Eq.* (5.12) *is*

$$\hat{f}_t(Q;\kappa,\gamma,m) = a_t^{\gamma-2}(m_Q)a_{con}(m_Q)\left(\sum_{i=1}^{\rho}\lambda_i \min\left\{u_i,\vec{u}_i\right\} + \lambda_{\rho+1}\vec{u}_{\rho+1}\right)\;, \tag{5.13}$$

where $\vec{u}_i \coloneqq \max\left\{0, 1-\sum_{j=1}^{i-1} u_j\right\}$ *and*

$$u_i \coloneqq \frac{\left(\max\left\{\mathbf{e}_{:m}{}^{\top}\mathbf{v}_{i\uparrow[Q]}, \mathbf{e}_{:m}{}^{\top}\mathbf{v}_{i\downarrow[Q]}\right\} - \frac{m}{n}\mathbf{e}^{\top}\mathbf{v}_i\right)^2}{a_{con}(m)}\;. \tag{5.14}$$

We can now compute a bound $\hat{f}_t(Q;\kappa,\gamma)$ over all subsets $R \subseteq Q$ using Lemma 5.2 as follows.

$$\hat{f}_t(Q;\kappa,\gamma) = \max_{m\in\{0,\ldots,m_Q\}} \hat{f}_t(Q;\kappa,\gamma,m) \tag{5.15}$$

$$\geq \max_{m\in\{0,\ldots,m_Q\}} \max_{R\subseteq Q,\,|R|=m} J(R) \geq J(Q)\;. \tag{5.16}$$

In the supplementary material we provide an algorithm that uses this approach to compute an upper bound in $O(n\rho)$ time, which is linear when ρ is considered a small constant.

5.3 Hyperparameter Optimisation

We now present methods to tune the hyper-parameters of the kernel to be used for the similarity assessment of the entities in our dataset. This process must be carried out in such a way that preserves important information for the task of anomaly detection or clustering, as introduced in Section 5.2.2.

However, there is no explicitly defined target variable available for classification or regression, and hence it is not possible to use standard schemes from supervised learning for this sub-task. Instead, we are given a set of predicates P, each of which can be seen as a classification variable. We therefore make two key assumptions in the derivation of our methods: 1) *the attributes of the datasets and thereby the predicates derived from them are relevant to the task for which the dataset was created*, and 2) *two subsets whose predicate description is similar should themselves be similar*. We therefore seek a method which takes into consideration all predicates at the same time without favouring just a single one or a few of them, and that admits a meaningful interpretation of predicate conjunctions.

5.3.1 Measuring the Fitness of a Candidate Kernel

We now consolidate these assumptions into a method to assess the fitness of a candidate kernel $\kappa : E \times E \to \mathbb{R}$.

The first obstacle is that we require to evaluate the performance of a kernel over entities, using ground truth over predicates. An straightforward way to induce a similarity over entities using their predicates is by using the intentions of each entity as a feature.

$$\kappa_{\text{lp}}(\epsilon_1, \epsilon_2) \coloneqq \frac{|\operatorname{int}(\epsilon_1) \cap \operatorname{int}(\epsilon_2)|}{\sqrt{|\operatorname{int}(\epsilon_1)||\operatorname{int}(\epsilon_2)|}}, \tag{5.17}$$

This is equivalent to the normalised intersection kernel, or the linear kernel over the characteristic vectors of the entity intentions. In addition to the downsides of the linear kernel, this kernel loses discriminating power when the number of predicates is small, and increases the complexity of the method to that of the number of entities, typically larger than the set of predicates.

Instead, our method relies on three key steps: first, we capture the similarity of each predicate to the others using a Tanimoto kernel [Tan58], which operates on the predicate set P. Then, we 2) use a kernel over sets to compute the candidate kernel, whose domain is the set of entities E, to define one over P. Finally, 3) we assess the fitness of the candidate kernel, as the alignment of the previously derived kernel and the Tanimoto kernel.

This process yields a fitness value for any candidate kernel on E, which can subsequently be used as a proxy by a method to either select one kernel out a family thereof, or to create a composite kernel.

The Tanimoto kernel [Tan58] is the usual name of the Jaccard Index when used as a kernel, and in its original form operates on the power-set of a ground set $\kappa_{\text{Tan}} : 2^E \times 2^E \to [0, 1]$, here on the space consisting of all entity subsets. We can therefore apply it on the set of predicates P through the use of the extension operator

$$\kappa_{\text{Tan}}(p_1, p_2) \coloneqq \frac{|\operatorname{ext}(p_1) \cap \operatorname{ext}(p_2)|}{|\operatorname{ext}(p_1) \cup \operatorname{ext}(p_2)|}, \qquad p_1, p_2 \in P\,. \tag{5.18}$$

The Tanimoto kernel is known to be positive definite [Gow71] and captures the normalised amount of shared structure between sets, which makes it a natural choice to assess predicate conjunctions. We can therefore use κ_{Tan} to measure the similarity of P. Owing to the assumption of meaningful selection of predicates, we hence treat their similarity as ground truth.

The candidate kernel κ operates on two entities. In order to compare it with the ground truth we need to transform it into a kernel that operates on sets of entities. Arguably the most appropriate choice is the kernel mean map [MFS+17] of κ. This is in turn also a positive definite kernel, and is based on the same assumptions used in Section 5.2.1 to derive the MMD: that each set contains i.i.d. samples of a distribution, which is mapped in the Hilbert space to the mean of the mappings of each element in the set. This is exactly the kernel that induces the MMD distance[4], and therefore fits naturally to our assumptions. When this kernel is applied on any two predicates $p_1, p_2 \in P$ it becomes

$$\kappa_{\text{mm}}(p_1, p_2; \kappa) = \frac{1}{|\operatorname{ext}(p_1)||\operatorname{ext}(p_2)|} \sum_{\substack{\epsilon_1 \in \operatorname{ext}(p_1) \\ \epsilon_2 \in \operatorname{ext}(p_2)}} \kappa(\epsilon_1, \epsilon_2)\,. \tag{5.19}$$

We now need a means to compare the ground truth similarity of predicates to the one induced by the candidate kernel κ through the kernel mean map embedding over the same predicates. For this we use an established similarity measure of two kernels, the kernel alignment [CSE+02]

$$\operatorname{algn}(K_1, K_2) \coloneqq \frac{\langle K_1, K_2 \rangle_{\text{F}}}{\sqrt{\|K_1\|_{\text{F}} \cdot \|K_2\|_{\text{F}}}}, \tag{5.20}$$

[4] To see this, note that the distance induced by a kernel is $d(p_1, p_2) \coloneqq \kappa(p_1, p_1) + \kappa(p_2, p_2) - 2\kappa(p_1, p_2)$, which can be verified to match Eq. (5.6).

where K_1, K_2 are the Gramians of the two kernels and $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ the Frobenius inner product[5]. This measure resembles the cosine similarity of the two Gramians with respect to the Frobenius inner product, and it is $0 \leq \operatorname{algn}(K_1, K_2) \leq 1$, where the lower bound is due to the definiteness of the Gramians. The upper arises from the Cauchy-Schwarz inequality, and therefore $\operatorname{algn}(K_1, K_2) = 1 \iff K_1 \propto K_2$.

Combining all the above, we define the **kernel fitness** of κ

$$\operatorname{kernfit}(\kappa; P) \coloneqq \operatorname{algn}\Big(\mathbf{K}_{\mathrm{Tan}}(P), \mathbf{K}_{\mathrm{mm}}(P; \kappa)\Big)\,, \tag{5.21}$$

where $P \subset 2^E$ is a set of predicates, $\mathbf{K}_{\mathrm{Tan}}(P)$ is the Gramian of the Tanimoto kernel over P, and $\mathbf{K}_{\mathrm{mm}}(P, \kappa)$ is the Gramian of the kernel mean map $\kappa_{\mathrm{mm}}(p_1, p_2; \kappa)$ for each $p_1, p_2 \in P$ and with κ the candidate kernel over entities.

We can now use the kernel fitness defined in Eq. (5.21) to select the best out of a family of kernels on E, for instance by an appropriate global optimisation scheme, such as grid-search or—as in our experiments—Bayesian optimisation.

5.3.2 Multiple Kernel Learning

A special case of measuring kernel fitness arises when the family of kernels we evaluate is a (positive) linear combination of a collection of constituent kernels. In this case, there exists a non-negative vector of coefficients $\alpha \in \mathbb{R}^{[}+]p$ such that the candidate kernel can be written as $\kappa_\alpha \coloneqq \sum_{\iota=1}^{p} \alpha_\iota \kappa_\iota$; then the (squared) kernel fitness of Eq. (5.21) becomes

$$\operatorname{kernfit}^2(\kappa_\alpha) = \frac{1}{\|\mathbf{K}_{\mathrm{Tan}}\|_{\mathrm{F}}^2} \frac{\alpha^\top \mathbf{v}\mathbf{v}^\top \alpha}{\alpha^\top \mathbf{W} \alpha}\,, \tag{5.22}$$

where $\mathbf{W} \in \mathbb{R}^{p \times p}$ with $\mathbf{W} \succeq 0$ and $\mathbf{v} \in \mathbb{R}^p$, defined as

$$\begin{aligned} W_{i,j} &\coloneqq \langle \mathbf{K}_{\mathrm{mm}}(\kappa_i), \mathbf{K}_{\mathrm{mm}}(\kappa_j) \rangle_{\mathrm{F}} \\ v_i &\coloneqq \langle \mathbf{K}_{\mathrm{mm}}(\kappa_i), \mathbf{K}_{\mathrm{Tan}} \rangle_{\mathrm{F}} \end{aligned}\,, \qquad i, j = 1, \ldots, p\,. \tag{5.23}$$

When the components κ_ι are guaranteed to be orthogonal Cristianini et al. [CSE+02] provides an optimal solution for α, which amounts to using a vector of coefficients whose elements are proportional to the alignment of each component. In practice, however, our candidate kernels can be not only non-orthogonal, but also highly correlated, which makes the $\mathbf{W}$ matrix

[5] It is $\langle K_1, K_2 \rangle_{\mathrm{F}} = \operatorname{Tr}\big[{K_1}^\top K_2\big]$ and $\|K\|_{\mathrm{F}}^2 = \sum_{i,j} K_{ij}^2$.

badly conditioned or even non-invertible. For these cases we modify the the solution that is optimal in the orthogonal case, into what forms the following heuristic.

We keep adding the components, considering them in decreasing order of their alignment. At each step, the added components are weighted with coefficients which are proportional to their fitness, which can be shown to be $\text{kernfit}(\kappa_\iota) \propto {}^{(\alpha_\iota v_\iota)^2}/_{W_{\iota,\iota}}$. In the end we pick the top-most coefficients from the beginning until the index that maximises the kernel fitness. Although, when orthogonal components are added the resulting alignment can only increase [CSE+02], adding a component that is correlated with an already added one may lower the resulting fitness. Thus, our selection scheme results in a sparse selection in case of highly correlated components, while it remains optimal in case of orthogonal ones.

5.4 Related Work

This method can be seen as the kernelisation of typical subgroup discovery. Indeed, Eq. (5.7) can be seen as a generalisation of the GWI of Eq. (2.10), where the generality term is replaced with a general scaling function a_t and the exceptionality term with the MMD measure, which contains the former as the special case in which the chosen kernel is the trivial linear one.

This makes our method also related to exceptional model mining (EMM), which, as we recall from Section 1.2, assesses the exceptionality of the subgroup as the difference between the coefficients of two fitted models, one on the dataset and the other on the subgroup. These models either require a well defined relation between the target variables (e.g., measuring correlation, using tests, etc), or one target must function as a classification label, so that a linear model can be trained over the rest of the variables. In contrast, is not limited by this, since we use the MMD statistic that can directly detect differences in the entity distributions. In addition, in EMM a new model must be estimated at each iteration, which can be rather costly for large problems, whereas in our case we simply need to pre-compute the kernel Gramian, after which the computation per iteration succeeds in linear time. Perhaps more importantly, the EMM methods do not provide an optimistic estimate and are therefore not able to efficiently perform exact optimisation. In fact, the paradigm of EMM could also benefit from the use of kernels, but to the best of our knowledge, this case has not yet been studied.

For the optimisation of our task our general problem cannot be expressed as a standard mathematical program, due to the dependency on the scaling function a_t, and we therefore use our IDDFS algorithm (see Section 2.4.2).

What can be computed as an integer quadratic program is the sub-problem Eq. (5.12) in computing our optimistic estimator, which is only restricted by a cardinality constraint. As we show in Appendix B.2.2, however, this computation is prohibitively inefficient, and in fact up to several orders of magnitude slower than our full optimisation pipeline.

We note that, due to its resemblance with the Rayleigh quotient, our problem also seems at a first glance relevant to its maximisation and minimisation schemes, by simply inverting the fraction. However, although several methods can solve the unconstrained Rayleigh quotient [Kon80; LSL12], our cardinality constraint makes them non-applicable to our problem. Also note that, contrary to continuous optimisation, it is not easy to first solve for the transformation z and then solve for the x in the unit box. Indeed, these methods rely on the particular structure of the 0-1 box, which is violated by the transformation we require. In fact, exactly because of the arbitrary scaling function a, generic bounds are not applicable, even the known naïve ones [Sho87]. A simple method could sort the values of the matrix, but our proposed bound is tighter and more computationally efficient, which makes a comparison equivalent to creating a straw man to later defeat.

When it comes to hyper-parameter optimisation, several methods have been proposed for un-supervised tuning parameters, which can be used for kernel clustering [LMA+16], or for general clustering [Mei18]. These methods however often require multiple clusters instead of just two, and also ignore the predicate information.

5.5 Experiments

We implement and evaluate our method on real world datasets, and here we demonstrate the results.

5.5.1 Datasets

Despite the abundance of structured datasets (e.g., containing images, graphs, time-series, etc) and similarly many with tabular data, there is a substantial scarcity of datasets with both such information at the same time. We thus compile three datasets that come close to practical tasks from drug discovery, finance and social sciences, and demonstrate related aspects from our hyper-parameter optimisation methods in Section 5.3.1. We next quickly outline the nature of these datasets, and we delegate detailed parameters to Appendix B.2.

In `Chem` we compile a dataset of drug-like molecules based on the *ChEMBL* [GBB+12]

database. These molecules constitute substances with potential pharmaceutical usage and their predicates are derived from suggestive pharmaceutical traits [MHA+10] annotated by human specialists, while their structure is assessed with a pre-computed kernel available through the *PubChem* interface [CTP10]. In `Stock` we describe stocks of companies listed in the New York Stack Exchange with indicative financial traits of each company, alongside a time-series of daily prices; these are used to assess stock similarity through extracted Rocket features [DPW20]. Finally, `Twitter` contains Twitter ego nets [LK14]: small subgraphs of the interaction network centered around selected individuals. Their attributes are followed users and used hash-tags. We compare their graphs using the state-of-the-art Wasserstein-Weisfeiler-Lehman kernel [TGL+19].

5.5.2 Kernel Hyperparameter Tuning

Except for the `Chem` dataset, which comes with a pre-computed kernel provided by the PubChem interface, we need to specify hyper-parameters for the kernels of our datasets. We therefore demonstrate here the methods introduced in Section 5.3.1, for both settings of single parameter and multiple kernel learning.

For the `Twitter` dataset we use the state-of-the-art WWL [TGL+19] kernel, κ_{wwl}, which requires the specification of a single scalar parameter γ_{wwl}. We choose this parameter by optimising the kernel fitness of Eq. (5.21) using Bayesian optimisation with a Gaussian process prior [SLA12] evaluated at 120 points (Fig. 5.3). For `Stock` each of the 1000 extracted Rocket features [GBB+12] yields a radial basis kernel whose σ parameters are individually tuned with the same procedure, resulting in a collection of equally many candidates for multiple kernel learning, which are highly correlated. To then combine these features into a single kernel we use the algorithm described in Section 5.3.2, which results in a sparse combination of only 4 sub-kernels (Fig. 5.4a).

As a measure of fitness for this method we also compare the average recall of the classification of each predicate as a classification variable, using a kernel with a parameter trained at each point. We show that the kernels chosen with kernfit yield significantly higher scores than when maximising the alignment of the linear predicate kernel in Fig. 5.3b for a broad range of γ_{wwl} values, and for the optimal multiple kernel coefficients arising from the two methods in Fig. 5.4b.

5.5.3 Necessity of Constrained Optimisation

We further motivate our method by demonstrating the necessity of constrained optimisation. Since unconstrained clustering is description-unaware it is extremely unlikely to yield a describable subset in the first place. Finding for it the closest description may also result in a low-quality subset. In Fig. 5.5 we show the centroids of the optimal named subset Q, a local optimum Q_{km} found by kernel k-means initialised with Q, and the closest named Q_{Jac} in the Jaccard sense to Q_{km}. Unsurprisingly Q_{km} scores the highest in our objective but has no description, while the naïvely found named one has 5 times lower quality than our optimum.

5.5.4 Efficiency of Computation

To optimise Problem 3 we use our branch and bound variant, for which a key factor deciding its efficiency is the ability of the optimistic estimator to prune the search space. That means that a key measure of this efficiency for a given dataset is the number of states it visits. Since there are no baseline optimistic estimators for our novel objective, as a comparable measurement we show (Fig. 5.6) the percentage of visited states over those required by an exhaustive search. Our optimistic estimator remains practically efficient even for higher rank matrices. Also, since we can only expressed our objective as an integer quadratic problem (IQP) for a fixed cardinality (see Eq. (5.12)), the full problem would require solving $O(n)$ hard IQP sub-problems. As we show in the extended version of this work, our method is superior to the IQP approach.

5.5.5 Discovered Subgroups

A selection of discovered subgroups is listed in Table 5.1, while the points of their subsets are depicted in Fig. 5.2 alongside the first two eigenvectors.

Quantitatively, we notice that the size of the discovered subsets is controlled by the tuning parameter γ, with a corresponding lowering of the measure of dissimilarity in $\mathcal{H}$.

Of special interest is the occurrence of the predicate [`sector` = *energy*] in several top subsets, indicating that the stock prices in this sector were the most deviating from the rest. Since we chose the price sequence in the data to cover the years of the pandemic, they highly overlap the period of heavy restrictions imposed in transport, which has been extensively reported to financially impede this sector.

5.6 Discussion

In our derivation of the objective function we relied on the maximum mean discrepancy, which is a statistic able to discern the differences between any two distributions when the kernel is characteristic, at least asymptotically. However, one should be careful not to misinterpret or overlook an important subtlety: that the discovered subgroup is not necessarily the one that is the most significant, a concept that we already addressed in Section 2.3. Instead, it is the optimal with respect to the specified balance of generality and exceptionality, which does, under certain cases[6], coincide with the most statistically significant one.

On the same note, efficient characteristic kernels do not exist for certain structures, for instance graphs. Indeed, having a polynomial characteristic kernel would amount to being able to solve the subgraph isomorphism problem, which is known to be NP-hard. Nevertheless, this is rarely a practical hindrance, as is evidenced by the great success of such kernels in general machine learning tasks.

When it comes to the complexity of our objective function itself, it is $O(nk)$, where k is the rank of the Gramian. This makes our method stand out due to its worst case quadratic computational complexity, whereas typical objectives in subgroup discovery require linear time. In practice, however, it is often the case that the Gramian is low-rank, for instance when the linear kernel is applied on low-rank data; otherwise, it is always possible to use a low-rank approximation of the Gramian, for instance using an appropriate Nystrom approximation.

5.7 Conclusion

In this chapter we provide all key components for a method that is able to find subgroups from datasets of entities with virtually arbitrary structural information. To achieve this we first propose an objective function that is based on the maximum mean discrepancy statistic, appropriately adapted for the subgroup discovery setting.

The proposed objective function employs a positive definite kernel to assess the similarity of each entity, and altogether provides the means to detect the exceptionality of a subgroup, while it also allowing to specify the importance of the subgroup generality.

[6]This happens when the structure of the entities is a Gaussian-distributed scalar, the linear positive definite kernel is chosen, and we select a weight of $\gamma = 1$. Then, our objective becomes optimisation-wise equivalent to the absolute of the z-score.

Since the peculiarity of our problem does not permit a direct use of standard solvers, we use our IDDFS algorithm, for which we develop an efficient optimistic estimator, that becomes tight when the Gramian is close to rank-1. We show that a tight optimistic estimator can also be formulated as a series of integer quadratic programs, but demonstrate that our approach is up to orders of magnitude faster than this alternative.

To complete the practical applicability of our method, we also propose a novel measure of kernel fitness, which we use to tune the hyper-parameters of candidate kernels. Our study also includes the setting of multiple kernel learning, for which we also provide a heuristic method that yields sparse sub-kernel combinations.

Importantly, we present results that show both that our method is practical and that its discovered subgroups are meaningful to a human user.

γ	Subset Description	$\lvert Q\rvert$	MMD
$[0.00 - 0.50]$	$[49 \leq \texttt{price}] \wedge [\texttt{sector} = \textit{Energy}] \wedge [10 \leq \texttt{mktCap}]$	0.005	0.2767
$[0.50 - 0.60]$	$[1.9 \leq \texttt{lastDiv}] \wedge [\texttt{sector} = \textit{Energy}] \wedge [9.8 \leq \texttt{mktCap}]$	0.008	0.2315
$[0.60 - 0.90]$	$[\texttt{sector} = \textit{Energy}] \wedge [\texttt{activelyTrading}] \wedge [4.8 \leq \texttt{volAvg}]$	0.074	0.0548
$[0.90 - 1.00]$	$[\texttt{sector} = \textit{Energy}] \wedge [\texttt{activelyTrading}]$	0.082	0.0500
$[1.00 - 1.25]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}] \wedge [0.00017 \leq \texttt{lastDiv}]$	0.529	0.0064
$[1.25 - 3.60]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}]$	0.691	0.0045
$[3.60 - 8.00]$	$[10 \leq \texttt{price}]$	0.834	0.0023

Table 5.1: Selected subsets from `Stock` and their metrics.

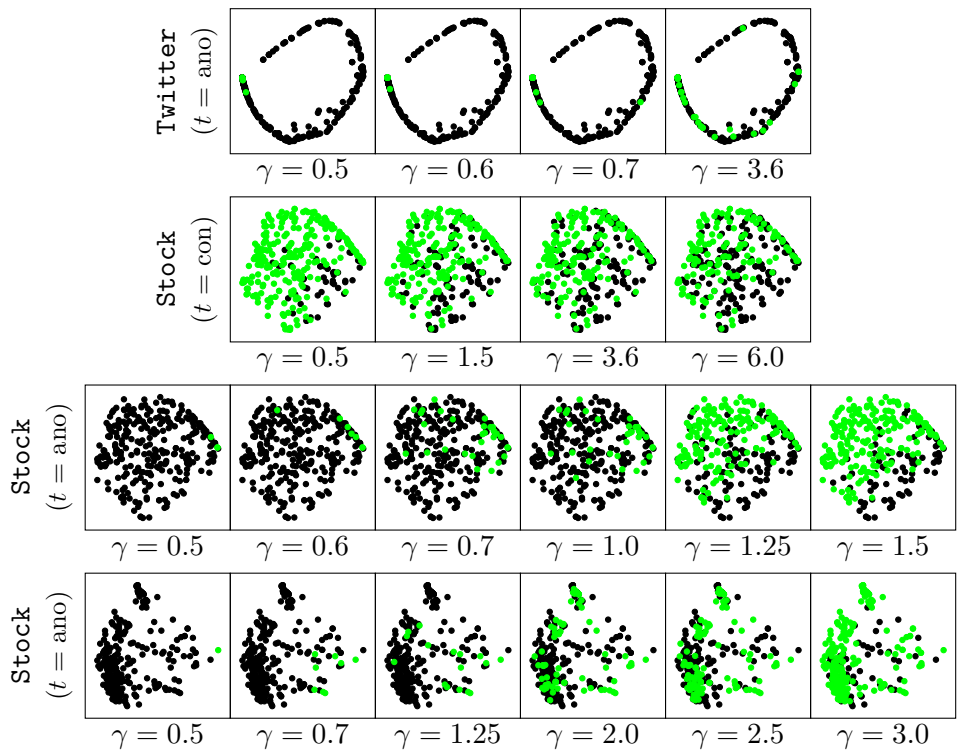
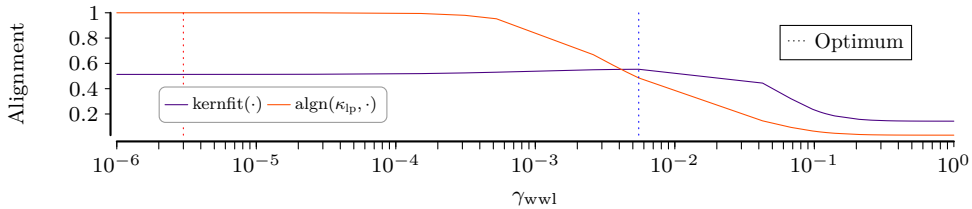
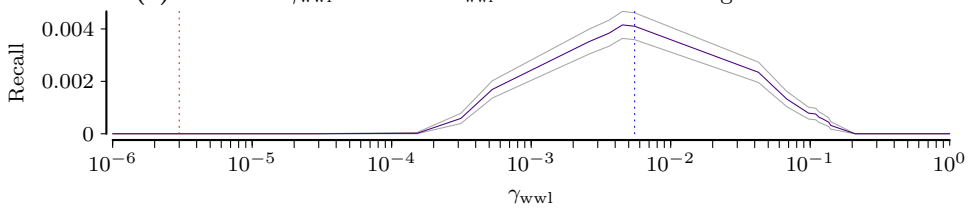


Figure 5.2: Selected discovered named subsets.

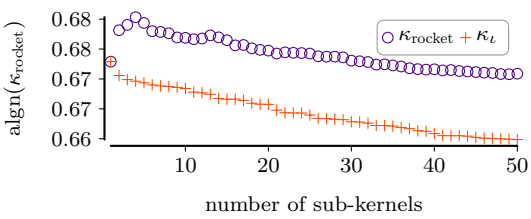


(a) Values of γ_{wwl} for which κ_{wwl} was tested for its alignments.

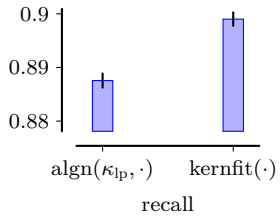


(b) Classification recall of predicates using $\kappa_{\text{wwl}}(\cdot)$ in a SVM.

Figure 5.3: Tuning the scalar parameter γ_{wwl} for `Twitter`, tested at 100 points selected through Bayesian optimisation [above], and the recall of predicate validities using the kernel for the given γ_{wwl} (average of 50 splits per predicate) [below].



(a) Optimal MKL coefficients.



(b) Classification score.

Figure 5.4: Multiple kernel learning for `Stock`: top-ranking sub-kernels are added until the resulting alignment stops increasing [left], and classification recall of the optimal kernel, compared against the naïve linear predicate kernel κ_{lp}.

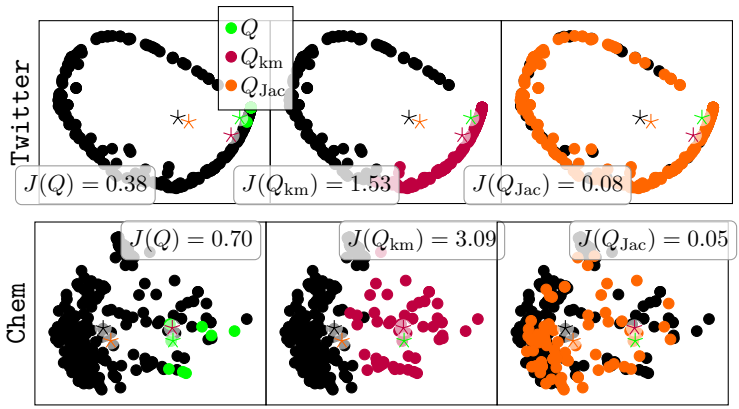


Figure 5.5: Naming the clusters of kernel k-means yields low quality subsets. Comparison of our discovered, optimal subset Q [left], the kmeans discovered subset without name Q_{km} [middle], and the closest named to the latter Q_{Jac} [right].

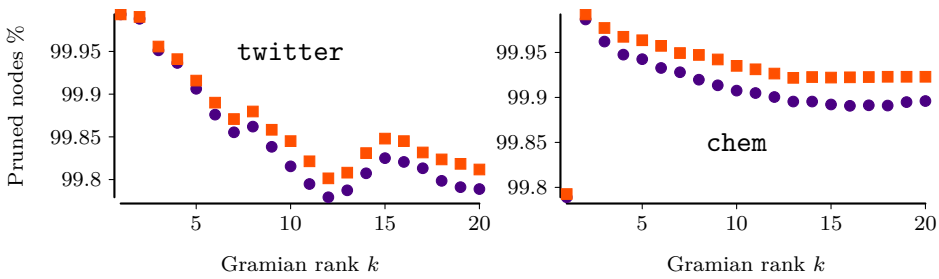


Figure 5.6: Efficiency of the optimisation in terms of visited search states during the branch and bound algorithm, as the matrix rank k of the Gramian increases.

6 A Structure-Aware Graph Kernel

> We wish to discuss a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biologic interest.
>
> *(Rosalind Franklin, Nature, 1953)*

Using the methods of Chapter 5 we can apply the powerful tool of subgroup discovery on any domain of entity target variables, simply by defining an appropriate positive definite kernel over this domain. In fact, this is only but a single application of positive definite kernels, as they can additionally extend any machine learning method that is based on inner products but also on distances, using the so-called *kernel trick* [Sch00]. This trick allows these standard tools to take into account the underlying structure between the data—for instance, whenever they lie in a low-dimensional manifold—but also to accommodate data whose entities are themselves complex structures. One of the most flexible and generally useful domains with structure that can accommodate the dataset entities is the domain of graphs, and we hence focus on the particular case of positive definite kernels defined over this very domain.

Graph kernels are known for almost two decades, and are notorious for their high requirement of computational resources. In this chapter we show how we can not only improve the flexibility, but in certain cases also the efficiency of a popular kind of graph kernels. Our kernel, similar to most graph kernels, can be expressed as an R-convolution [Hau99], which can be seen as an extension of the notion of convolution that operates on kernels defined on sub-structures of these graphs, and in a way that preserves the positive-definiteness of those kernels. Our kernel belongs to the family of random walks kernels, which have been among the first such formulations [GFW03], and through their more recent generalisations [VSK+10] they still remain relevant for certain configurations [KJM20].

Intuitively, given an alignment ϕ between the vertices of one graph to those of another, one can count the number of simultaneous random walks between the two graphs. These are random walks performed on two graphs, starting from one vertex in one graph and one vertex in the other; then

they advance the vertex of each graph one edge at a time in lockstep, so that at each step the vertices of two graphs respect the given alignment ϕ. Of course, in general such an alignment is not available. The random walk kernels avoid the dependency on any one particular mapping between the vertices of two graphs by instead considering all simultaneous walks over all possible alignments, between the vertices of one graph and those of the other.

The number of all such walks can be elegantly formulated in terms of linear-algebraic operations, thus allowing the use of a variety of algorithms for its computation, while being very flexible: they allow to incorporate arbitrary distributions over the vertices of the graphs as starting and stopping points of the walks, different weights for different walk lengths, node and edge similarities encoded as kernels, as well as the possibility to incorporate edge labels [VSK+10]. This allows, for example, to down-weigh vertex alignments in the random walks between vertices that we know (or strongly suspect) to be dissimilar.

Recent work [NML+18; TGL+19; SHW+21] suggests that (non-random-walk) kernels that align vertices based on structural properties improve predictive performance on several datasets. Examples of such properties are the coreness of vertices, vertex degree, or Weisfeiler-Lehman labels. In these cases, vertices with the same or *similar* value share a similar structure and should arguably be aligned in a random walk kernel using a suitable kernel on structural property values.

In this work, we additionally focus on the case that very dissimilar vertices are not just down–weighed, but are not allowed to be simultaneously visited by a random walk at all. Assuming that a good estimation of vertex similarity was available then one would expect that the count of walks that respect such vertex alignments would preserve or improve the accuracy of the walk, while the other walks might be considered noise. In many cases this results in more fine grained similarity functions that give high similarity to graphs that have many simultaneous random walks that respect structure, e.g., simultaneously travelling through similarly dense regions.

Consider the two small graphs G, H in Fig. 6.1. Assume we know that vertices with label 1 are dissimilar to label 3 vertices, but those with label 2 are somewhat similar to both. The vanilla random walk kernel would consider the full product graph in Fig. 6.1c, modelling vertex alignment using a kernel function on vertices. On the other hand, the actual information that is relevant in this scenario is represented by the much smaller product graph of Fig. 6.1b, while considering only identically labelled vertices as matches would arguably lose too much information, as shown in Fig. 6.1a.

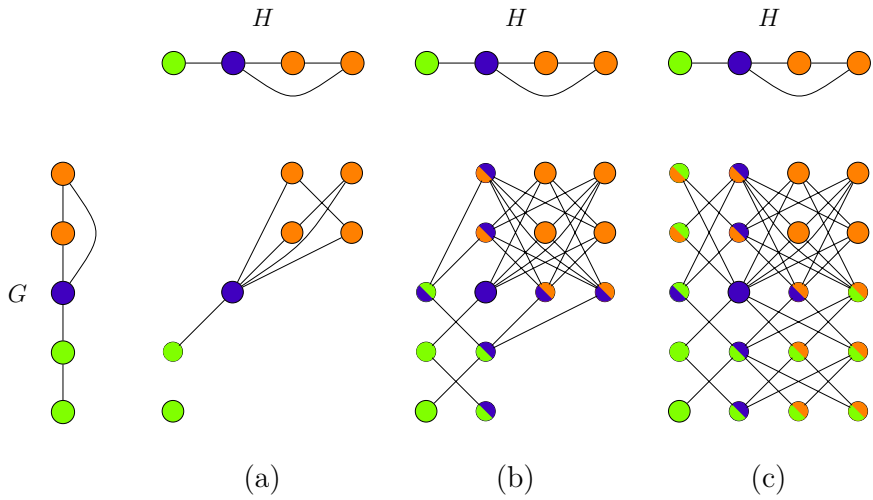


Figure 6.1: Two small graphs G, H with structural labels 1, 2, 3 indicated as colours. Considering only vertex pairs of identical labels as in (a) arguably [GFW03] results in too sparse product graphs. The product graph (c) considered by the generic random walk kernel [VSK+10] includes all pairs of vertices. Our proposed kernel may disallow the mapping of label 1 to 3 and can then be computed on (b).

We thus capture this assumption as a restriction of the allowed alignment during the simultaneous walk, and encode this as a vertex kernel of bounded support on structure-aware ordinal labels, which can be extracted, for instance, from structural properties of the vertices. For the resulting random walk kernel, we present a computational method that is not only asymptotically, but also practically faster than alternative computational methods. This superiority increases as we enforce stricter assumptions and only allow alignments of closer labels. In practice, this restriction formally corresponds to using a vertex kernel with a smaller support. If (almost) complete alignment information is available, such as, for example in brain activity networks [SJX+17], our method can be computed in up to quadratic time, in contrast to the cubic time of the generalised framework. This further improves accordingly when using sparse graph representations.

Our contributions of this chapter are as follows:

- We propose a fast algorithm to compute random walk kernels based on bounded support vertex kernels.
- We thus study the gap between the rather restrictive choice of Gärtner et al. [GFW03] and the general framework of Vishwanathan et al. [VSK+10].
- We describe a class of bounded support vertex kernels for integer-valued

structural attributes.
- We study the performance of several such attributes when used within our framework.
- We empirically show significant improvements on some datasets, while generally being on par with the vanilla random walk kernel.

Although we present this work as a stand-alone kernel, perhaps a more important goal is to show that even for datasets on which we do not outperform the vanilla kernel in terms of accuracy, we still attain shorter running time. Additionally, our work can also improve state-of-the-art kernels that use random walks as a component [NV20].

6.1 Preliminaries

We begin with an overview of the necessary concepts that we make use of in our analysis and introduce the relevant notation. For conciseness we first define $[n] \coloneqq \{1, \dots, n\}$ to be the set of the first n positive naturals; we then, additionally, define $[n]_0 \coloneqq \{0, 1, \dots, n\}$.

In this work we consider undirected graphs $G = (V, E)$. The set of edges can also be represented as the **adjacency matrix** $\mathbf{A} \in \mathbb{R}^{n \times n}$, which has entries $[\mathbf{A}]_{i,j}$ equal to 1 if $(v_i, v_j) \in E$, or 0 otherwise, where we assume a fixed ordering $v_1, \dots, v_n$ of V. To describe edge-weighted graphs we straightforwardly replace the adjacency matrix with its weighted version. Since G is undirected, the adjacency matrix is symmetric: $\mathbf{A} = \mathbf{A}^\top$. For two matrices $\mathbf{A}' \in \mathbb{R}^{n' \times m'}$ and $\mathbf{A}'' \in \mathbb{R}^{n'' \times m''}$, we write their Kronecker product as $\mathbf{A}' \otimes \mathbf{A}'' \in \mathbb{R}^{n'n'' \times m'm''}$; similarly, for $\mathbf{A}', \mathbf{A}'' \in \mathbb{R}^{n \times m}$ we denote their Hadamard product as $\mathbf{A}' \circ \mathbf{A}'' \in \mathbb{R}^{n \times m}$. In the following analysis we generally consider a pair of graphs, G' and G'', in which case we implicitly refer to property p of the first graph as p' and of the second as p''.

6.1.1 Counting Simultaneous Walks over Aligned Vertices

The basic random walk kernel applied on two graphs $G' = (V', E')$ and $G'' = (V'', E'')$ is equal to the number of simultaneous walks between them. More precisely, assume a mapping $\phi : V' \to V''$ between the nodes of the two graphs; then, a simultaneous walk would only be allowed to traverse an edge $e = (u', v') \in E'$ only whenever an edge also exists in G'' between the mapped vertices of it $(\phi_{u'}, \phi_{v'}) \in E''$. Equivalently, we could visualise such a walk as one over the graph G_ϕ with vertices $(u', \phi_{u'})$ for all $u' \in V'$ and edges $E_\phi \coloneqq \{((u, v), (\phi_{u'}, \phi_{v'})) \mid (u', v') \in E' \wedge (\phi_{u'}, \phi_{v'}) \in E''\}$.

The random walk kernel avoids the dependency on any one particular

such mapping ϕ. To achieve this, it considers all possible mappings by performing a simultaneous walk on the graph with vertices the Cartesian product $V_\times \coloneqq V' \times V''$ and edges the union over all possible vertex mappings

$$E_\times \coloneqq \bigcup_\phi E_\phi = \{((u', u''), (v', v'')) \mid (u', v') \in E' \wedge (u'', v'') \in E''\} \; ; \quad (6.1)$$

this configuration results in the *direct product graph* $G_\times \coloneqq (V_\times, E_\times)$. Importantly, a walk on $G_\times$ is equivalent to a simultaneous walk on graphs G' and G''. Indeed, consider advancing a walk on $G_\times$ from node $(u', u'') \in V_\times$: we can interpret this as first randomly selecting a mapping ϕ which i) respects the current node $\phi_{u'} = u''$, and for which ii) E_ϕ contains at least one edge with an endpoint (u', u''); then traversing one of these edges from E_ϕ at random. As a small parenthetical remark, in this work we propose to down-weigh—or outright exclude—certain of these alignments ϕ.

6.1.2 The Random Walk Kernel

The adjacency matrix $\mathbf{A}_\times$ of $G_\times$ is equal to the Kronecker product of the adjacency matrices of G' and G'' [Wei62], $\mathbf{A}_\times = \mathbf{A}' \otimes \mathbf{A}''$. Additionally, the i-th power of the adjacency matrix $\mathbf{A}^i$ of a graph contains in its element $[\mathbf{A}^i]_{i,j}$ the number of walks with exactly i steps from the i-th to the j-th vertex of the graph. Hence, the number of all such walks can be written as $\mathbf{e}^\top \mathbf{A}_\times^i \mathbf{e}$, where $\mathbf{e} = (1, \ldots, 1)^\top$ is the vector of all ones. The random walk kernel that counts all simultaneous random walks is thus defined as

$$k(G', G'') = \sum_{i=0}^{\infty} \gamma_i \mathbf{e}^\top \mathbf{A}_\times^i \mathbf{e} \, , \quad (6.2)$$

where the constants γ_i ensure the convergence of the series [GFW03].

Among different choices for the sequence γ, two special cases arise, which allow the analytic computation of the series in Eq. (6.2). The geometric sequence $\gamma_{geom}(i) \coloneqq \lambda^i$ defines the **geometric** random walk kernel, while the power series coefficients of the exponential function $\gamma_{exp}(i) \coloneqq 1/i!$ define the **exponential** one. Using symbolic forms these are

$$\begin{aligned} k_{geom}(G', G'') &\coloneqq \mathbf{e}^\top (I - \lambda \mathbf{A}_\times)^{-1} \mathbf{e} \text{ and} \\ k_{exp}(G', G'') &\coloneqq \mathbf{e}^\top \exp(\lambda \mathbf{A}_\times) \, \mathbf{e} \, , \end{aligned} \quad (6.3)$$

where, in the former, λ takes any value small enough so that $\|\lambda \mathbf{A}_\times\| < 1$, as is necessary for the geometric series to converge, while in the latter, λ serves as a positive parameter.

The random walk kernel of Eq. (6.2) as proposed by Gärtner et al. [GFW03] has been extended or adapted in several ways. Borgwardt et al. [BOS+05] use a kernel on vertices and edges to define the edge similarities of the direct product graph. Additionally, the random walk can be equipped with starting/ending probabilities and weighted matrices. These can all expressed as instances of the general random walk formulation [VSK+10]

$$k(G', G'') = \sum_{i=0}^{\infty} \gamma(i) \mathbf{q}^\top \mathbf{W}_\times^i \mathbf{p} \,, \tag{6.4}$$

where $\mathbf{p} = \mathbf{p}' \otimes \mathbf{p}''$ and $\mathbf{q} = \mathbf{q}' \otimes \mathbf{q}''$ are start and stop probabilities on vertices and $\mathbf{W}_\times$ contains generalised edge similarities, as computed by a kernel between the edges of the two graphs. These generalisations can now be incorporated in the analytic expressions of Eq. (6.3); this gives

$$\begin{aligned} k_{geom}(G', G'') &\coloneqq \mathbf{q}^\top (\mathbf{I} - \lambda \mathbf{W}_\times)^{-1} \mathbf{p} \text{ and} \\ k_{exp}(G', G'') &\coloneqq \mathbf{q}^\top \exp(\lambda \mathbf{W}_\times)\, \mathbf{p} \,, \end{aligned} \tag{6.5}$$

where the parameter λ plays a similar role as in Eq. (6.3). Now, the formulation of Eqs. (6.4) and (6.5) allows for a vertex kernel that encodes the similarity of vertex combinations in the generalised edge similarities.

6.1.3 Computing the Random Walk Kernel

In theory, the geometric and exponential kernels of Eq. (6.2) can be computed by solving a linear system and computing a matrix exponential, respectively, each of which operate on a matrix that is easily derived from $\mathbf{A}_\times = \mathbf{A}' \otimes \mathbf{A}''$. In practice, however, both these operations are cubic in general, and since $\mathbf{W}_\times \in \mathbb{R}^{n'n'' \times n'n''}$, they would require an excessive computation of $O(n^6)$.

To this end Vishwanathan et al. [VBS06] note that the spectral decomposition of $\mathbf{A}_\times$ can be efficiently computed as $S_\times D_\times S_\times^\top = (S' \otimes S'')(D' \otimes D'')(S' \otimes S'')^\top$. This allows one to compute Eq. (6.3) as $\mathbf{q}^\top S_\times \left(\sum_{i=0}^{i_{\max}} \gamma(i) D_\times^i \right) S_\times^\top \mathbf{p}$, where the—otherwise cubic—operation need only operate on the diagonal matrix $D_\times$. As an additional benefit, the spectral decomposition of each adjacency matrix may only be computed once for all graph pairs; then the entire Gram matrix needs only $O(n^2 m^2 i_{\max} + mn^3)$ time for m graphs.

Alternatively, the same work used the Conjugate Gradient iterative solver to compute k_{geom} of Eq. (6.3). Each iteration performs a matrix-vector operation with a Kronecker matrix by employing the identity

$$(\mathbf{A}' \otimes \mathbf{A}'')\mathbf{x} = \mathrm{vec}(\mathbf{A}'' \,\mathrm{mat}(\mathbf{x}) \mathbf{A}'^\top) \,, \tag{6.6}$$

where the operation vec : $\mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ creates a vector by concatenating all columns of a matrix and mat : $\mathbb{R}^{mn} \to \mathbb{R}^{m\times n}$ is the inverse operation. Using this identity within iterative methods requires $O\left(n^3k\right)$ computations per kernel entry, assuming k iterations.

However, when a vertex or an edge kernel is used, as in the general framework of Eq. (6.4), the matrix $\mathbf{W}_\times$ does not have a Kronecker decomposition. Then, these and other proposed methods are not applicable any more, and iterative methods must be used, which are based on repeated matrix-vector operations. These methods also allow using an edge kernel as long as it has a known feature representation on a finite d-dimensional Hilbert space [VSK+10]; the matrix-vector operation is then performed on each of the d dimensions of feature vectors, which needs $O\left(n^3d\right)$ computations.

As a workaround, the resulting similarity matrix can be approximated by its nearest Kronecker product, although not without downsides: the additional complexity of a rank-1 singular value decomposition on the full matrix [Loa00], as well as the potential of arbitrary quality degradation due to the involved approximation.

6.1.4 Graph Concepts

As usually, $N(v) \coloneqq \{u \in E \mid (u,v) \in E\}$ are the neighbours of a vertex v and $\delta(v) \coloneqq |N(v)|$ its **degree**. The **induced subgraph** $G[U] \coloneqq (U, \{(u,v) \in E \mid u,v \in U\})$ of a vertex subset $U \subseteq V$ is the subgraph of G with vertices U and those edges of E with both endpoints in U. Similarly to Section 4.1, we also denote $\delta_U(u)$ the degree of vertex u in the induced subgraph $G[U]$.

In our kernel we will also make use of the concept of graph coreness, the structural property of vertices that was introduced in Section 4.1. We recall that a **k-core component** of a graph G is an (inclusion-wise) maximal connected subgraph of G whose vertices U have all a degree of at least $\delta_U(u) \geq k$. The subgraph comprising all k-core components of this graph is its **k-core** $H(k)$, with vertices the set $E(k)$.

The example of a core decomposition of Fig. 4.2 shows that the k-cores are nested to form a hierarchy over the vertices, so that the **k-shell** of G is the set of vertices that lie in the k-core but not in the $k+1$-core (same-coloured vertices in the figure). In this way, the k-shells define a partitioning over the vertices: the **core decomposition** of G, $\kappa_G : V \to [K]_0$, which assigns to each vector v its **core number** (or **coreness**) $\kappa_G(v) \coloneqq \{k \mid v \in E(k)\}$. Importantly, the core value of each vertex in the graph is a structural property of this vertex, that is subject to stronger structural constraints than its degree, due to the additional requirement that we count edges that must belong to the same k-core.

6.2 Structure-Aware Vertex Similarities

We now introduce the main object of our study: a random walk kernel that uses a vertex kernel based on structural attributes of the vertices, namely their degree or coreness. This vertex kernel can be combined with any other vertex kernel and/or an edge kernel, thus remaining fully flexibile.

The reason underlying our use of structural graph properties is twofold. First and foremost, very frequently, the available graphs come without any vertex label information. Even when vertex labels are available, however, they might either not offer sufficient flexibility on their own, or be unsuitable as the basis of assessing the alignment of random walks. For instance, when the vertex labels are noisy, the number of random walks can vary greatly, especially when the "hub"-vertices of one graph are mislabelled. Additionally, often times the available labels are too coarse, as is the case of binary variables. At other times, the labels might not admit a sensible way to compare them, beyond the simple test for their equality, as is typical for labels of categorical nature. In all these cases, simply allowing the alignment of exclusively the vertices with the same label might turn out to be too restrictive, and greater flexibility would be desired, instead. Therefore, the available vertex labels can be enhanced (or simply replaced) with derived structural vertex properties, such as vertex degree or vertex coreness; these properties are typically not subject to labelling noise and allow for a natural way to compare vertices with unequal but similar values, since a similar value also indicates a similar structure. Of particular interest is the case of the vertex coreness, which is efficient to compute and indicates a notion of "*robust connectedness*" in a graph [Sei83], as we demonstrated in Chapter 4. Using vertex coreness or similar values as vertex labels can further enables the use of a multitude of available kernels on scalars for the evaluation of vertex similarity. Choosing a smoother amongst them allows for greater resilience against edge or label noise,[1] while retaining the discriminating potential of these labels, which offers a good middle ground between using a too restrictive or too loose vertex alignment.

6.2.1 The Structural Similarity Random Walk Kernel

The use of such derived values for this purpose is an extension that can be incorporated in the random walk kernel of Eq. (6.4) by using as edge weight matrix $\mathbf{W}_\times$ the values of a particular edge kernel $k_E : V'^2 \times V''^2 \to \mathbb{R}$ on any pair of edges (u', v') and (u'', v'') with $u', v' \in V'$ and $u'', v'' \in V''$. First, let

[1] There are some graph classes and perturbation models for which this seems necessary if the structural similarity is defined by coreness [AV13].

us use $n' = |V'|$. Using the above trick, we can now express this operation by defining the edge similarity matrix to be

$$[\mathbf{W}_\times]_{(i-1)n'+r,(j-1)n'+s} = k_E\big((v'_i, v'_j), (v''_r, v''_s)\big)\,. \tag{6.7}$$

Here, the edge similarity kernel k_E can be decomposed into two constituents,

$$k_E\big((u', v'), (u'', v'')\big) \coloneqq k_{\mathrm{adj}}\big((u', v'), (u'', v'')\big) k_{struc}(u', u'') k_{struc}(v', v'')\,, \tag{6.8}$$

where k_{adj} is a kernel on graph edges and k_{struc} the structurally-aware vertex similarity kernel. In place of k_{adj} we can use any kernel, as long as it satisfies the constraints of Section 6.1.3: that is, one that must have a known and low-dimensional representation. Nevertheless, for the sake of notational simplicity, during our analysis we adopt the linear kernel on the graph adjacency entries; then the Gram matrix of this kernel becomes equal to the Kronecker product of the two adjacency matrices[2]

$$k_{\mathrm{adj}}\big((v'_i, v'_j), (v''_r, v''_s)\big) \coloneqq \big[\mathbf{A}'\big]_{i,j} \big[\mathbf{A}''\big]_{r,s}\ , \tag{6.9}$$

$$K_{\mathrm{adj}} = \mathbf{A}' \otimes \mathbf{A}''\,. \tag{6.10}$$

Without loss of generality, we can consider the structural vertex kernel k_{struc} as a kernel k_{att} over the image of a feature map $l_G : V \to \mathcal{X}$ that extracts **structure-aware properties** from the vertices of each graph, i.e.,

$$k_{struc}(v', v'') \coloneqq k_{att}\big(l_{G'}(v')\,, l_{G''}(v'')\big)\,. \tag{6.11}$$

As its Gramian goes over all pairs $V' \times V''$, it has a rank of 1, and we can express it using only a vector $\mathbf{k} \in \mathbb{R}^{n'n''}$:

$$K_{struc} = \mathbf{k}\mathbf{k}^\top, \text{ with} \tag{6.12}$$

$$\mathbf{k}_{(i-1)n'-r} \coloneqq k_{struc}(v'_i, v''_r)\,. \tag{6.13}$$

Finally, the edge similarity matrix can be written in vectorised form as

$$\mathbf{W}_\times = (\mathbf{A}' \otimes \mathbf{A}'') \circ K_{struc}\,. \tag{6.14}$$

[2]The Gramian of any edge kernel k_{adj} that satisfies the mentioned constrains can be expressed (or approximated) as the sum of several Kronecker products similar to Eq. (6.10), with one term for every dimension of the Hilbert space representation of the kernel. With little added effort, the following theory can be extended, albeit with an accompanying increase in computational complexity.

One important result of the above observations is that, together, they allow us to adapt the identity of Eq. (6.6) to accommodate the special structure required by our kernel.

Lemma 6.1 (Matrix-Vector Operation)**.** *The matrix-vector operation of the edge similarity matrix of Eq.* (6.14) *can be computed in* $O(n'^2n'' + n'n''^2)$ *as*

$$\mathbf{W}_\times \mathbf{x} = \mathrm{vec}\Big[\mathbf{T} \circ \big(\mathbf{A}''(\mathbf{T} \circ \mathrm{mat}[\mathbf{x}]\,)\mathbf{A}''^\top\big)\Big]\ , \tag{6.15}$$

where $\mathbf{T} \coloneqq \mathrm{mat}[\mathbf{k}]$ *is the matricisation of Eq.* (6.13)*.*

— For the proof see Appendix A.4.

For greater insight, we note that the matrix $\mathbf{T}$ can be regarded as the lower off-diagonal block of the vertex kernel Gramian, as applied on the concatenation of the vertices of the two graphs $[v'_1, \ldots, v'_{n'}, v''_1, \ldots, v''_{n''}]$; that is, we can define its entries as

$$[\mathbf{T}]_{i,j} = k_{struc}(v''_i, v'_j) = k_{att}\big(l_{G''}(v''_i)\,, l_{G'}\Big(v'_j\Big)\,\big)\ . \tag{6.16}$$

A closer look into this formulation reveals that, when there are elements of $\mathbf{T}$ for which the involved vertices are not considered similar according to k_{struc}, this matrix becomes sparse; then, we can exploit the structure of the zero entries of this matrix to avoid substantial parts of the matrix-vector computations $\mathbf{W}_\times \cdot \mathbf{x}$ that have no effect on the final result, yielding a more efficient computation.

Corollary 6.2. *Let at most* τ *pairs* (v', v'') *have a non-zero similarity according to* k_{struc}*, i.e.,* $|\{(v', v'') \in V' \times V'' \mid k_{struc}(v', v'') \neq 0\}| = \tau \leq n'n''$*. Then the matrix-vector operation of Eq.* (6.15) *can be computed in* $O\Big((n' + n'')\,\tau\Big)$*.*

A special instance of the above arises when the structural attributes extracted from the graph vertices lie on a set of *discrete scalars*, and since l_G can be an arbitrary function, we can assume without any loss of generality that $l_G \in \{1, \ldots, L\} =: [L]$. Particular interest lies in the case when we use an integer kernel k_{att} that is of **bounded support**, i.e., when there exists a $\delta \geq 0$ such that $k_{att}(i, j) = 0$ for all $|i - j| > \delta$. We refer to this threshold δ as the kernel **bandwidth**.

We hence call **Structural Similarity random walk (SUSAN)** the random walk kernel that uses as vertex kernel k_{struc} a bounded support kernel k_{att}, the latter of which is defined over integer-valued structural attributes that are derived from the graph vertices—typically the vertex coreness. In

this case, the resulting vertex similarity kernel can be expressed as

$$k_{struc}(v', v'') = k_{att}\Big(l_{G'}(v')\,, l_{G''}(v'')\Big) k_{\mathrm{v}}(v', v'')\,, \tag{6.17}$$

where k_{v} is either a pre-existing vertex kernel to be incorporated, or can be simply assumed to be the constant function and be therefore omitted. The resulting kernel remains both positive-definite and of bounded support. Due to its last property, the kernel considers only potentially beneficial relations, while gaining special structure that enables a significantly more efficient computation.

6.2.2 Avoiding Inconsequential Calculations

We now study the useful special structure of SUSAN and analyse its computational complexity. First, and without loss of generality, we assume the rows and columns of each adjacency matrix to be arranged so that they correspond to vertices in increasing order of $l_G(v)$. This gives rise to the block representation of the adjacency matrix $\mathbf{A}$ of graph G as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \cdots & \mathbf{A}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{L,1} & \cdots & \mathbf{A}_{L,L} \end{bmatrix}\,, \tag{6.18}$$

where each block $\mathbf{A}_{i,j}$ contains the (possibly empty) block of all edges (u, v) from vertices with $l_G(u) = i$ to those with $l_G(v) = j$, and L is the maximal value of l_G. It will be of help to define the size of each block as $\mathbf{A}_{i,j} \in \mathbb{R}^{b_i \times b_j}$, where $b_i \coloneqq |\{u \in V \mid l_G(v) = i\}|$. We can now compute the exact number of non-zero elements τ of $\mathbf{T}$, by observing that this matrix becomes a banded block matrix with block-bandwidth δ, and whose each block has equal elements.

Lemma 6.3 (Operation Complexity)**.** *The matrix-vector operation $\mathbf{W}_{\times}\mathbf{x}$ for the SUSAN kernel whose vertex kernel has a bounded support with bandwidth δ can be computed in exactly*

$$c = (2n'+2n''+1)\sum_{i=0}^{K'} b'_i \sum_{j=\max(0,\lambda-\delta)}^{\min(K'',i+\delta)} b''_j - n'n'' \tag{6.19}$$

floating point operations, which is $O\big((\delta+1)(n'+n'')B^2\big)$, for B the maximal size among all b'_k, b''_k. In contrast, a naïve computation requires $2n'n''(n'+n'')$ such operations.

— For the proof see Appendix A.4.

When we use the SUSAN kernel equipped with coreness values as vertex, we can complete Lemma 6.3 by providing a lower bound for the calculated computational complexity. This is based on a worst-case instance of a graph that attains the largest possible block size B defining the size of a block, at least in the case of coreness.

Lemma 6.4 (Block Size Lower Bound)**.** *Let G be a simple graph on n vertices. Then there exists a k-shell with at least $\sqrt{n}$ vertices.*

Proof. Suppose there are at most b vertices in any k-shell of G. Note that the κ_G-shell is equal to the κ_G-core of G. Hence the κ_G-core of G contains at most k vertices. Any k-core must contain at least $k+1$ vertices, hence, $\kappa_G \leq k-1$. Assuming that all b-shells for $b \leq \kappa_G$ are nonempty, this implies that there are at most $k \cdot k$ vertices in G. Thus, there must exist at least one k-shell with at least $\sqrt{n}$ vertices in G. □

Hence, according to Lemma 6.3, no better bound than $O\left(\delta(n'+n'')n\right) = O\left(\delta n^2\right)$, where $n = \max(n', n'')$ can be given for the SUSAN kernel.

6.2.3 Selection of the Structural Attribute

Although within the context of the SUSAN kernel we can plug in any integral structural attribute, the greatest efficiency comes with small bandwidths. For this reason, two natural such properties arise as particularly helpful: the vertex degree and its coreness, which can both be computed at virtually no cost.

As a side-note, the choice of the vertex kernel may also affect the matrix-vector complexity c. By using the vertex degree, a bound on the non-zero entries of each adjacency matrix row also arises, given the degree that each block belongs to. Combined with the bound for the non-zeros of the matrix $\mathbf{T}$, we can replace the innermost sum in Eq. (6.19) with $\sum_{j=\max(0,i-\delta)}^{\min(L'',i+\delta)} \min(i, b_i)$. In the case of coreness the quantity c can be bounded from below as $c \in o((n'+n'')\sqrt{n'}\sqrt{n''})$, even with a bandwidth of $\delta = 0$, since there always exists a k-shell with at least $\sqrt{n'}$ vertices in any simple graph G' (see Lemma 6.4).

6.2.4 Selection of the Attribute Kernel

Note that, in order to ensure positive definiteness for the SUSAN kernel, all its constituents must also be positive definite, and therefore also k_{att}. We

hence define a class of bounded support positive definite kernels that can be used in the role of k_{att}, to assess the similarity of integer-valued structural vertex properties.

To fully specify k_{att} as a positive definite kernel it is sufficient to define its feature mapping ϕ on some Hilbert space. For a given bandwidth δ and an arbitrary **shape vector** $s \in \mathbb{R}^{\lfloor \delta/2 \rfloor + 1}$ we can define such a feature map $\phi_s^\delta : \mathbb{Z} \to \mathbb{R}^\omega$, whose image contains (countably) infinite-dimensional elements with indices $\lambda = \ldots, -3/2, -1, -1/2, 0, 1/2, 1, 3/2, \ldots$:

$$\phi_s^\delta(i)_\lambda \coloneqq \begin{cases} s_{\lceil |i-\lambda| \rceil} & 2|i-\lambda| \leq \delta, \delta \text{ odd}, 2\lambda \text{ odd} \\ s_{|i-\lambda|+1} & 2|i-\lambda| \leq \delta, \delta \text{ even}, \lambda \in \mathbb{Z} \\ 0 & \text{otherwise} . \end{cases} \tag{6.20}$$

The image of ϕ_s^δ is a (countably) infinite-dimensional vector space that becomes Hilbert through the natural inner product. Thus, a positive definite kernel k_{att} over $\mathbb{Z}$ can be defined as

$$k_{\text{att}}(i,j) \coloneqq \langle \phi_s^\delta(i) \, , \phi_s^\delta(j) \rangle_{\mathbb{R}^\omega} \coloneqq \sum_{\lambda=-\infty}^{\infty} \phi_s^\delta(i)_\lambda \cdot \phi_s^\delta(j)_\lambda \ , \tag{6.21}$$

and can be easily verified that it has bounded support with bandwidth δ. Additionally, it also belongs to the class of **shift invariant kernels**; i.e., its value only depends on the difference of its entries: $k_{att}(i,j) = k_{att}(|i-j|)$, where we slightly abused notation. Since we assume that the structural properties are captured as an integer by the structural labels l_G, shift-invariance is a natural property that avoids making any further assumptions on these structural properties.

Note that not all shift-invariant bounded support functions are positive definite, but any function in the form of Eq. (6.21) is. Out of these, arguably the simplest one arises when s is the constant vector with elements $s_i = \frac{1}{\sqrt{\delta+1}}$. This choice yields the kernel whose graph resembles a triangle

$$k_{att}(i,j) = \max\left(0, 1 - \frac{|i-j|}{\delta+1}\right) , \tag{6.22}$$

and is also the one we adopt to complete Eq. (6.17).

6.2.5 Computation of SUSAN

Among the contributions of this work is a proof-of-concept implementation of the main BLAS3 components of the matrix-vector operation $\mathbf{W}_\times \mathbf{x}$ of the SUSAN. This implementation uses the key observation formalised in

Lemma 6.3, and therefore has a complexity that is upper bounded by that of its naïve computation.

We recall from Section 6.1.3 that fast algorithms are available for the vanilla random walk kernel. However, when using a vertex and/or edge kernels, like in the case of SUSAN, these methods cannot be applied, since the similarity matrix $\mathbf{W}_\times$ does not have a Kronecker decomposition. Then iterative methods have to be used, all of which rely on an efficient computation of the matrix-vector operation $\mathbf{W}_\times \mathbf{x}$.

We complete the work of [VSK+10] by applying the iterative method of Al Mohy et al. [AH11] to compute the exponential version of SUSAN. This method involves a truncated Taylor expansion, in which the order is computed for the required accuracy based on a bound on the norms of the matrix. We empirically study the convergence of this method in Section 6.3.1.

With this we establish that both practical algorithms for the computation of the SUSAN kernel benefit from a more efficient implementation of this matrix-vector operation.

Note that an additional advantage of our algorithm is that it only needs to store the τ elements of the $\mathbf{x}$ vector. This not only improves the cache locality of the data during computation, especially in the case of small δ, but can also improve the convergence of the used solver (c.f., Section 6.3.1).

6.3 Experiments

We now evaluate our proposed algorithm to compute the random walk kernel for bounded-support vertex kernels, as well as the utility of several structural similarity measures within this framework. In particular, we consider coreness, vertex degrees, and application specific structural vertex similarities. SUSAN is implemented[3] in C++ and Python.

As an example of data with application specific vertex similarities, we use three brain connectome datasets [SJX+17]: these represent connections between brain regions (as vertices) that are consistently labelled by integers across patients. Furthermore, close-by labels indicate functional proximity[4]. These, as well as the other datasets considered below are publicly available at Kersting et al. [KKM+16].

[3] Available at https://eda.mmci.uni-saarland.de/prj/susan/

[4] Vertices with similar labels have a much higher than expected probability of being connected by an edge in these datasets, which in turn implies a high correlation between the EEG time series of the two regions. We observe (cf. Section 6.3.2) that a nontrivial bandwidth on the labels increases the kernel performance.

6.3.1 Efficiency

To evaluate the theoretical advantage of our implementation on real-world datasets, we first compare our algorithm for computing SUSAN against a random walk kernel with a vertex kernel of unbounded support. To this end, we compute the vanilla random walk kernel with an appropriate iterative method for each kernel kind (see Section 6.2.5), as the faster methods of [VSK+10] are then inapplicable. We hence obtain an implementation-independent measurement by comparing the time required to compute SUSAN using the same implementation for both our algorithm and the baseline matrix-vector operation. For simplicity, and since the matrices of these datasets are of small dimension, we assume full matrix storage. We do note, however, that highly optimised implementations of the baseline matrix–vector operation (MVO) could outperform our proof-of-concept algorithm in certain hardware, despite its theoretical superiority.

On each brain connectome dataset, and for an increasing bandwidth parameter δ, we use the iterative schemes of Section 6.2.5 to compute both the exponential and the geometric SUSAN kernels. In the top row of Fig. 6.2 we compare the elapsed wall-clock time when using our proof-of-concept implementation of the matrix-vector product against the baseline computation to compute the same kernel. We see that for both kernels and for small bandwidths a speedup of up to an order of magnitude is attained; with increasing δ the runtime slows down to that of the naïve algorithm.

Although the key contributing factor in the more efficient computation is a faster MVO computation, we also study the required number of iterations as a potential secondary factor of efficiency. Therefore, in the bottom row of Fig. 6.2 we compare the average number of MVOs for SUSAN against those for the vanilla random walk kernel, using the same iterative method as above. We observe that the conjugate gradient solver (geometric variant) converges faster for the lower bandwidth vertex kernels, primarily due to the lower dimension of the problem in the case of SUSAN. Since, however, the difference in required MVOs is smoother than the running time, there seems to be an additional factor in play. We conjecture this to be the low-pass property of the vertex kernel of SUSAN, which seems to impose more smoothness in the Krylov Space that this solver uses.

6.3.2 Accuracy

Next we investigate the predictive performance of bounded bandwidth structural vertex similarities combined with random walk kernels. We show that using a bounded bandwidth kernel does not deteriorate performance com-

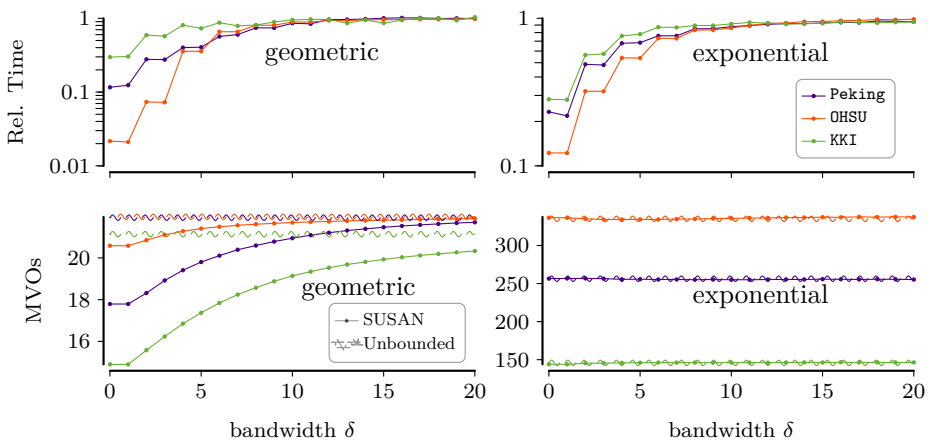


Figure 6.2: [Top] Relative time performance of SUSAN vs. random walk with unbounded vertex kernel (100%). [Bottom] Absolute number of matrix-vector operations (MVOs) required for the iterative computation of SUSAN.

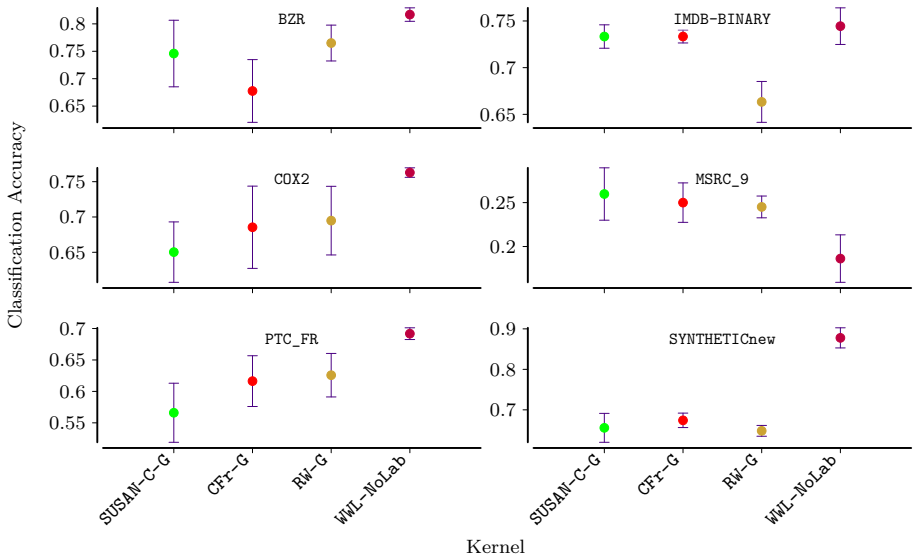


Figure 6.3 [Higher is Better]: Classification accuracy and standard deviations of SVM classifiers on benchmark datasets (geometric variants). Provided for the sake of completion, as these datasets do not yield rich enough structural vertex properties.

pared to the vanilla random walk kernel and other related state-of-the-art graph kernels. We instantiate the (exponential resp. geometric) random walk kernel with integer kernel on vertex labels as described in Section 6.2.4

using coreness (SUSAN-C-E, resp. SUSAN-C-G), vertex degree (SUSAN-Deg-E, SUSAN-Deg-G), and Weisfeiler-Lehman labels (SUSAN-WL-E, SUSAN-WL-G). We compare against vanilla random walk kernels (RW-E, RW-G), the Core Framework kernel (CFr-E, CFr-G) [NML+18], as an alternative to combine random walks with degeneracy, and against the Wasserstein-Weisfeiler-Lehman kernel [TGL+19] as a representative state-of-the-art alternative, without using vertex labels (WWL-NoLab).

We evaluate each candidate kernel by the accuracy of a C-SVM classifier [FCH+08] equipped with it. We compute random 80%/20% train/test splits, using stratified random sampling. On each training set we use a 3-fold cross validation to identify optimal hyper-parameters of each kernel using a grid search with 15 samples. For every random walk kernel we search the λ parameter in the range $\lambda \in [10^{-6}, 5]$ for the exponential and in $\lambda \in [10^{-5}, 1]$ for the geometric variants. For the WWL kernels we use the values $\lambda \in [10^{-5}, 10]$, as suggested by the authors. The SVM regularisation parameter is selected from $C \in [10^{-3}, 10^{5}]$. Both λ and C are sampled using a logarithmic scale. For the truncated versions of our kernels we use grid search over the set $\delta \in \{0, 1, 2, 3, 4, 5, 10, 15\}$.

For the sake of completeness, we first report results on standard graph kernel benchmark datasets in Fig. 6.3. Our kernel is never significantly worse than the Core Framework or the vanilla random walk kernel on these datasets. However, it is significantly better than the latter on the IMDB-Binary dataset ($p = 0.012$). This indicates that the computation can be sped up without hurting predictive performance.

Figure 6.4 reports results of our kernel variants and competitors on brain connectome graphs. It shows that our bounded bandwidth kernel with Weisfeiler-Lehman labels achieves highest predictive performance on average on OHSU, our kernel with coreness highest performance on Peking_1 and its geometric variant on KKI. To estimate the statistical significance of these results, we repeated each process from the beginning on a fresh split of each dataset, for 30 iterations and use Welch's two sample t-test with significance level $p = 0.05$. Table 6.1 shows the p-values of two tests, comparing SUSAN-C-G and SUSAN-C-E to the competitors. There is no statistically significant performance difference between the SUSAN variants and the vanilla random walks, except for the Weisfeiler-Lehman based kernels that perform (comparatively) well on OHSU, and poorly on Peking_1. When compared to the core framework kernel, SUSAN seems mostly on par with it, except on Peking_1 where CFr-E performs poorly in comparison, and KKI, where SUSAN-C-G is significantly better than its core-framework equivalent ($p = 0.0027$). Due to the increased computational effort of the core framework

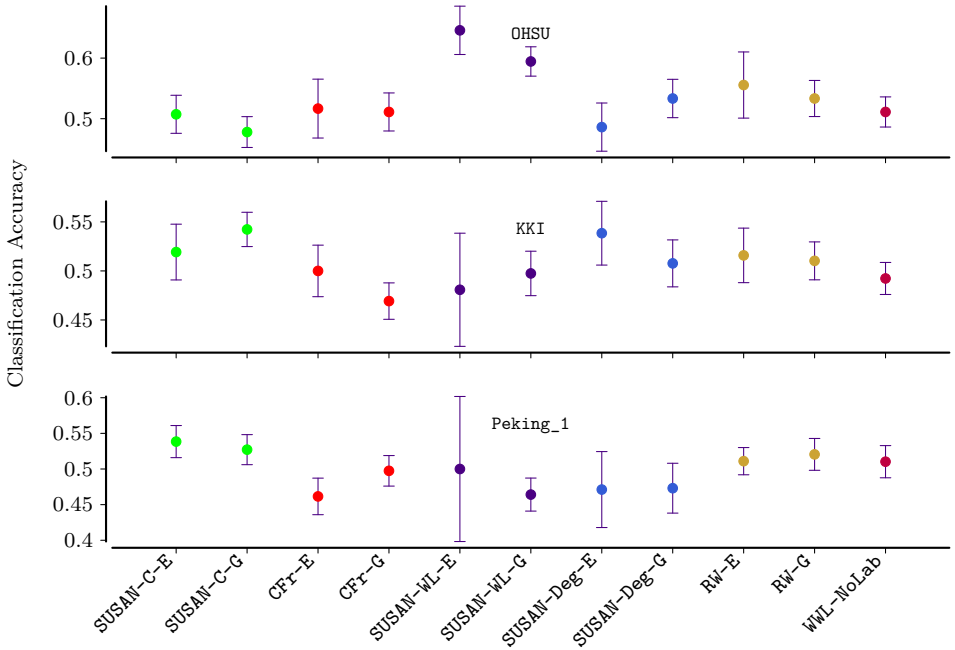


Figure 6.4 [Higher is Better]: Classification accuracy and standard deviations of SVM classifiers on brain connectome datasets.

(cf. Section 6.4) using our kernel and coreness is hence beneficial.

Finally, to assess the usefulness of degeneracy as a vertex label extractor in the case of no label information, we also compare it against the WWL kernel without label information (`WWL-NoLab`). We find that SUSAN performs well when compared with other methods without vertex information, and even significantly outperforms the state-of-the-art for the `KKI` (p-value=0.02). To the same end, we also report the results for the WWL kernel, in which we replace the default use of degrees with coreness (`WWL-Core`), with occasionally better performance.

For the sake of completeness we also compare against the full WWL (`WWL`). Unsurprisingly, `WWL` outperforms SUSAN, since the latter is not given access to label information.

6.4 Related Work

Gärtner et al. [GFW03] introduced Random Walk kernels, initially with a runtime of $O(n^6)$. Notably, their kernel already allows to incorporate

Dataset	SUSAN type	CFr-E	CFr-G	SUSAN-Deg-E	SUSAN-Deg-G	RW-E	RW-G	SUSAN-WL-E	SUSAN-WL-G	WWL	WWL-Core	WWL-NoLab
KKI	SUSAN-C-E	0.3115	0.0738	0.6694	0.3788	0.4781	0.3976	0.0525	0.2757	0.7940	0.3664	0.2080
	SUSAN-C-G	0.0983	**0.0027**	0.4594	0.1248	0.2148	0.1114	**0.0064**	0.0609	0.6112	0.0854	**0.0199**
OHSU	SUSAN-C-E	0.8820	0.5345	0.3420	0.7192	0.6919	0.7253	0.9045	0.9834	0.9930	0.8986	0.5382
	SUSAN-C-G	0.9598	0.7924	0.5669	0.9101	0.8603	0.9187	0.9745	0.9992	0.9998	0.9901	0.8226
Peking_1	SUSAN-C-E	**0.0111**	0.0690	0.1306	0.0517	0.1131	0.2623	0.3184	**0.0060**	0.9457	**0.0153**	0.1609
	SUSAN-C-G	**0.0302**	0.1633	0.1766	0.0973	0.2605	0.4148	0.4074	**0.0243**	0.9659	0.0592	0.2932
KKI	SUSAN-C-E	0.6885	0.9262	0.3306	0.6212	0.5219	0.6024	0.9475	0.7243	0.2060	0.6336	0.7920
	SUSAN-C-G	0.9017	0.9973	0.5406	0.8752	0.7852	0.8886	0.9936	0.9391	0.3888	0.9146	0.9801
OHSU	SUSAN-C-E	0.1180	0.4655	0.6580	0.2808	0.3081	0.2747	0.0955	**0.0166**	**0.0070**	0.1014	0.4618
	SUSAN-C-G	**0.0402**	0.2076	0.4331	0.0899	0.1397	0.0813	**0.0255**	**0.0008**	**0.0002**	**0.0099**	0.1774
Peking_1	SUSAN-C-E	0.9889	0.9310	0.8694	0.9483	0.8869	0.7377	0.6816	0.9940	0.0543	0.9847	0.8391
	SUSAN-C-G	0.9698	0.8367	0.8234	0.9027	0.7395	0.5852	0.5926	0.9757	**0.0341**	0.9408	0.7068

Table 6.1: Listing of p-values for the Welch t-test: comparing the performance of SUSAN (using coreness) against all other kernels on the brain connectome datasets. Null hypotheses: [top] SUSAN is not better than X, [bottom] SUSAN is not worse that X. That is, a **blue** value [top] indicates that SUSAN significantly outperforms X, whereas a **red** value [bottom] means SUSAN is significantly outperformed by X. We note that WWL has access to vertex labels, and thus gets an unfair advantage.

integral vertex labels in a way that corresponds to the trivial bandwidth $\delta = 0$ in our setting. Vishwanathan et al. [VBS06; VSK+10] proposed several iterative methods which reduced the computational complexity to $O(n^3)$. They generalized this to a framework for random walk kernels which can incorporate nonuniform starting and stopping probabilities on vertices, different weights for different walk lengths, node and edge similarities encoded as kernels, as well as the possibility to incorporate edge labels. However, their framework cannot benefit from bounded support vertex kernels. Kang et al. [KTS12] propose to speed up random walk graph kernel computation using a low rank approximation of the adjacency matrix $\mathbf{W}_\times$ of the product graph, which in turn can be computed implicitly by low rank approximations of $\mathbf{A}'$ and $\mathbf{A}''$. They obtain a runtime of $O(Kn^2r^4 + r^6 + mr)$ for K different labels, where r is the rank parameter. In contrast, our method is exact (up to numeric precision) and runs in $O(\delta nB^2)$ for bandwidth $\delta \leq K$ and $B \leq n$.

Structural vertex labels have been used before to define graph kernels. For a recent broad-scoped survey, see Kriege et al. [KJM20]; we focus here on the combination of random walk kernels and structural properties. Mahé et al. [MUA+04] first propose to use *Morgan indices* as vertex labels in random walk kernels. Nikolentzos et al. [NML+18] introduce a graph kernel framework based on core decompositions. They suggest to compute kernels of the form $k(G', G'') = \sum_{k=0}^{\infty} k_{\text{base}}(H(k)', H(k)'')$ explicitly, which increases the complexity of the kernel computations by a factor K. In this work, however, we use the structural information given by the core decomposition to speed up the kernel computations. We thus don't fall into their framework; notably, our kernel tends to get faster when many core values are present and allows more fine-grained control over vertex alignments.

6.5 Discussion

Among other contributions, this work focuses on the little studied domain within the family of random walk kernels with respect to the treatment of vertex labels, where existing methods either only consider identically labelled vertices [GFW03] or do not take into consideration vertex labels in order to remain efficient. This is due to the existence of a simple decomposition for the matrix $\mathbf{W}_\times$ of Eq. (6.15) in the above trivial cases. Hence, it remains an interesting research question whether a similar decomposition would be possible for the general case, despite the fact that currently no known such decomposition exists for the generic Hadamard transform, that we show is required twice in the general case.

For use within our framework we provide a shift invariant positive definite

kernel on integers that has bounded support. This raises the question of whether the structural difference between vertex labels can be reflected sufficiently solely between the difference of their values, without taking the actual values into consideration; in other words, how reasonable is it to assume that a 4-core as similar to a 7 core as it is to a 1-core.

For the needs of this work we provide an optimised implementation for the case of non-sparse matrix storage, which remains competitive for certain configurations even to highly optimised BLAS3 libraries. We note, however, that the proper use of the intrinsic peculiarities of each architecture is a grand technological feat that exceeds the scope of this work, as it requires often times vectorised assembly instructions specialised for each architecture. This complexity is even more pronounced in the case of sparse matrices, although theory predicts similar computational advantages.

6.6 Conclusion

In this chapter we proposed a random walk kernel for graphs with vertices equipped with labels, and in specific such labels that can be compared based on another kernel defined on the vertex labels. We use this kernel on labels to specify the amount of alignment allowed during the counting of the random walks, which completes the two known extreme cases, and for certain cases offers superior performance.

Additionally, we showed how we can compute our kernel using iterative solvers, both in the case of geometric and exponential variants, and we demonstrated experimentally the convergence rate in real world settings. In the special case where the vertex labels are integers, we showed that the main matrix-vector computation required by the iterative solver can be written in a compact block form. When the kernel defined on the vertex labels is of bounded support, we showed that this matrix-vector formulation can be significantly sped up by avoiding inconsequential computations and provided an exact computation of the required floating point operations. Therefore, under certain conditions on the distribution of vertex labels, we can achieve even a linear computational complexity (with respect to the graph vertex count), instead of the typical cubic one of the generic random walk kernel. This performance is predicted by Lemma 6.3 and also shown experimentally.

Additionally, for cases where vertex labels are not available, we proposed a scheme to extract structure-aware integer vertex labels based on the coreness of each graph vertex. This scheme makes it meaningful to not only compare their values, as small difference in label value indicates similar structure, but also to impose a threshold beyond which no alignment between the

involved vertices is allowed, thus resulting in a bounded support kernel on vertices. In fact, we also provide a meaningful integer positive definite kernel of bounded support that can be used for such labels, which altogether forms SUSAN, which is the main object of study, while its constituent theoretical contribution have significant impact also independently.

We demonstrate that SUSAN can be more efficient to compute in specific cases, but also that it exhibits competitive performance to significant improvements in terms of accuracy on benchmark datasets, even compared to other state-of-the-art graph kernels.

7 Summary and Conclusion

The primary premise of this thesis has been that important subgroups within the data convey valuable information to a human audience. Therefore, these subgroups necessitate a theoretical and practical framework to efficiently discover, which is embodied by subgroup discovery. This is a framework that complements powerful methods in machine learning, since the latter serve a very different goal: that of globally modelling data in order to automate decisions, with much less focus on providing intelligible results to a human. In contrast, we motivated the value to a human audience of providing intelligible descriptions for subgroups: interesting local sub-populations that are akin to "nuggets" in normally vast amounts of data. We demonstrated our claim by describing the sub-population which exhibited the most severe SARS-CoV-2 symptomatology out of studied British patients: the subgroup of the Bangladeshi minority in London. We discussed how this finding was corroborated by discovering a mutation in the Bangladeshi genealogy, and showcased the clear benefits that a describable sub-population can bring to human audiences, for example of how this aided the battle against the pathogen.

Having established the value of subgroup discovery, we proceeded with a thorough overview of typical methods in the literature that we later completed with a formal in-depth analysis. For this, we adopted a statistical perspective to classify the existing methods in a common landscape, even though most of the original works did not place themselves in the statistical landscape but instead often provided just an intuitive justification of key decisions in their methodology. More precisely, we re-interpreted the established impact function as a distance between two statistics: one over the subgroup and another over a sample that is used to estimate a Null model used for assessing the exceptionality of the subgroup; we then described the implication of each of the two optimal choices for this latter sample, as the one derived by the assumptions of either the i) goodness-of-fit problem or ii) the two-sample problem, and discussed when should each of the two approaches be used. Importantly, we extensively studied a generalisation of the impact function: the geometrically weighted impact (GWI) f_{gwi}, and brought attention to its two components: the generality and exceptionality.

Further, we mentioned that these two components appear in the objective

functions of existing methods in the literature, but in each case with a fixed relation between them. From this standpoint, we studied further the GWI, and showed that the optimal subgroups discovered for the values of its trade-off parameter $\gamma \in [0, 1]$ are Pareto optimal with respect to these two components. We also proved that by this method we can find all the Pareto optimal points that lie on the convex hull of the logarithmically transformed values of the two components of generality and exceptionality. This result essentially enables our optimisation algorithms to function as a multi-objective optimisation method that can provide a concise subset of the Pareto frontier; that is, this subset is larger than simply the points on the convex hull of the points on this frontier, without containing optima that are similar to their neighbours. To the best of our knowledge, our proposed method is not only the first work that uses this approach within the subgroup discovery framework, but also for general purpose multi-objective optimisation.

Additionally, we discussed the existing state-of-the-art optimisation frameworks for our problem, which culminated in the proposal of our own algorithm. We first studied standard mathematical programs, and were the first to describe the optimisation domain for our problem as a polytope embedded in the integer lattice. After a brief mention of constraint satisfaction problem solvers, we concluded with our dedicated iterative deepening variant of the branch and bound algorithm that is specialised for the different subgroup languages proposed in the literature.

During our overview of the existing subgroup discovery methods, we quickly pinpointed key limitations that restricted the applicability of these methods to data with no structure or at best very simply structured data. This motivated the main work of this thesis, which contributes a colleciton of methods applicable to increasingly more complex structured data.

We started this journey with the notion of representative subgroups, which are not only exceptional with respect to one target concept, but also representative with respect to a control variable. The implication of introducing the control variable were studied from two similar perspectives. On one hand, we showed how we can use it to control gross trends in the data. For this, we recall the case study of German elections where geography is correlated with voting behaviour; here, we demonstrated that using the region as a control is necessary to be able to describe sub-populations which affect both regions equally. From a similar perspective, we can also use our approach to control for sensitive attributes, as was the case of discovering the most violent sub-populations within a dataset of recorded crimes, however now regardless of biases due to victim sex or perpetrator race.

Next, we focused on entities that are not identically distributed, but instead come equipped with relational information between them, that can be encoded as graphs where the entities are vertices and the relational information is represented as undirected edges between these vertices. We then steered our attention toward finding the most well-connected sub-populations of such entities. In our attempt to define what a good measure of connectedness is, however, we demonstrated that despite being a popular choice, the commonly used edge-to-vertex ratio is not a good measure. Indeed, we showed that certain commonly occurring subgraphs—such as bipartite ones—may even be fully disconnected by removing only a few vertices, even though they can have a relatively high edge-to-vertex ratio. We therefore introduced the concept of robust connectedness that is based on the minimum number of vertices that need to be removed from a subgraph so that it becomes disconnected, and introduced a measure for this notion using the average vertex coreness. This resulting measure is not only very efficient to compute, requiring only linear time, but is also able to capture deeper structural relationships between vertices. We subsequently used this notion within the subgroup discovery framework to finding the most robustly connected sub-population of entities. Within this setting we demonstrated that it is necessary to compute the robust connectedness exclusively within the subgraph induced by the vertices corresponding to the subgroup entities, which we showed to prevents the use of existing subgroup discovery algorithms. We therefore provided a measure of robust-discovery that takes the candidate subgraph into account, which we use in the core of our proposed RoSi algorithm that finds the most robustly connected subgroup.

Our contributions toward finding subgroups over entities with structure culminated with the introduction of positive definite kernels to assess the similarity between the studied entities. Using this approach it becomes possible for the first time to find the most exceptional sub-population among entities with virtually any kind of structure, such as stocks, molecules, or graphs. To achieve this we introduced a kernel-aware objective function that was based on the established statistic of maximum mean discrepancy in the role of an exceptionality component, which is then appropriately scaled with a term that captures generality. Originating from a statistical perspective, we showed that our approach can easily be used for both the goodness-of-fit problem and the two sample problem, with the only difference boiling down to a change in the form of the scaling function. Importantly, our objective function contains as a special case the geometrically weighted impact function as a special case, which arises when we use the linear kernel over a scalar target variable; moreover, in this case our optimistic estimator becomes tight

and has the same complexity as the tight optimistic estimator for the GWI. Despite its generality, however, our approach depends on the choice of a good kernel. While in typical machine learning tasks there is a metric that can be used for the optimisation of the kernel hyper-parameters, our case differs in that there is no clear-cut classification or regression variable, or any established such metric to assess the fitness of the candidate kernels. To solve this problem we treated the derived predicates as the ground truth of our task, after which we introduced a novel measure of kernel fitness, using which we make possible to optimise for the kernel hyper-parameters with standard techniques of global optimisation. In addition, we also studied a multiple kernel learning paradigm for the case where multiple sub-kernels over the entities are available, and described the way to compute the optimal coefficients, whenever possible, while for the rest of the cases we provide a heuristic method that yields a sparse combination of sub-kernels.

An important takeaway of our kernelised approach is that it constitutes a standalone, generalised framework that works out-of-the-box for virtually any kind of entity structures. This comes in contrast to the standard procedure required previously to develop a subgroup discovery method for a specific type of data structure. This procedure required several steps of substantial difficulty, including the development of an optimisation algorithm, the specification of an objective function, and derivation of a matching bound. In contrast, our method of kernelised subgroup discovery simply requires the specification of a Gramian matrix from a user. In other words, our optimisation algorithm, objective function and optimistic estimator remain the same, regardless of the underlying structure of each entity.

Having paved the way to kernelised subgroup discovery, we then improved on existing state-of-the-art graph kernels by developing SUSAN: a graph kernel that belongs to the class of random walk kernels and allows for the fine-grained control of how vertices are aligned during the counting of the random walks. More precisely, we studied undirected graphs with scalar vertex labels, and completed existing works that either i) allow the alignment of only identically labelled vertices, or ii) disregard the labels altogether. Our kernel allows for the alignment of different vertices to be weighted based on a kernel between their label values; we also provided a natural choice for such a kernel between integer labels. One limitation of our work is that as we depart from the simpler, existing solutions, certain mathematical simplifications are not any more applicable; this leads to an increased computational complexity when using direct linear algebra solvers. However, we showed how to compute our kernel very efficiently by using iterative solvers, which remains the de-facto approach for large, sparse graphs,

even in the case of the aforementioned simpler variants. Additionally, we focused on the special case of graphs with integer-valued vertex labels, which arises, for instance, in the study of brain connectome graphs. For such graphs we showed that the core linear operator attains a block Toeplitz representation; importantly, this matrix becomes banded when the used kernel on the vertex labels has a compact support, and depends on the size of this support. In addition, we developed a natural such kernel for the case of integer labels, that allows a user-specified size of its support. Using these components, we showed how to exploit this block representation to significantly reduce the computational complexity, as we showed in an optimised implementation of this matrix-vector operation. Our code belongs to the class of BLAS2 operations, and an efficient implementation often requires specialised assembly instructions for each architecture. For this reason, we limited ourselves to dense matrix representations, that not only serves as a proof-of-concept, but is also practically efficient for a small bandwidth parameter of our kernel. Finally, we obviate the need for provided integer vertex labels, by deriving them as the coreness of each vertex, which not only indicate a natural, structural similarity between the similarly-labelled vertices, but is also very efficient to compute. Although SUSAN has been motivated by the main subject of this thesis, its applicability is by no means limited within the framework of subgroup discovery; in fact, we demonstrated a superior classification accuracy for several classes of graphs when compared to the state-of-the-art Wasserstein-Weissfeiler-Lehman kernel against a standardised set of benchmark datasets.

7.1 Outlook

Our work both introduces novel paradigms in the field of subgroup discovery and broadens the applicability of existing endeavours.

By proposing representative subgroups, we introduced the first approach to provide exceptional subgroups, which at the same time remain representative with regard to a control variable. When this control variable is a sensitive binary trait, such as gender or race, we thereby ensure that the resulting subgroups are not discriminatory against a sensitive sub-population, and we therefore obtain the first fairness-aware subgroup discovery method. Although our method can i in theory applicable to multiple classes, we currently lack an efficient algorithm for the computation of its tight optimistic estimator beyond balanced, binary cases, which are the natural next steps in this direction. With regard to fairness, our approach is currently limited to the notion of statistical parity. Many other measures of fairness have

been proposed, and it makes for engaging future work to investigate how subgroups that are fair in these senses could be discovered. Most notable such formulations are those of i) equal opportunity and ii) equal odds, both of which involve a careful assessment based on contingency matrices, similar to previous methods that employ the same mechanism for causal discovery [BBV21].

Beside intelligible descriptions, the optimality of a subgroup also conveys particular information. In our work we shortened the distance between the mathematical formulation of popular objective functions and the human audience that needs to intuitively interpret what it means for a subgroup to be optimal with respect to these objectives. For this, we revisited the components of these objective functions from the perspective of multi-objective optimisation and we formally describe the set of optima along the Pareto frontier of exceptionality and generality. An important next step that closes even more this distance, is to broaden the set of methods that can find the most significant subgroup with respect to a given Null hypothesis, which has a direct appeal to medical personnel and other experts that have a well-developed statistical intuition for hypothesis testing. Although we mentioned a given case that maximising the objective function yields the most significant subgroup (i.e., when we use GWI with a trade-off parameter of $\gamma = 1/2$ for uni-variate Gaussian targets), this is generally not the case. It is therefore an important endeavour to develop methods that can find the most significant subgroup under different distributions and statistical assumptions, such as those governing our method for kernelised subgroup discovery.

By proposing kernelised subgroup discovery we are the first to leverage the versatility and power of positive definite kernels from within subgroup discovery. Due to these desirable traits of positive definite kernels and the vast amount of research available on kernels, on the one hand, and the enormous application potential of subgroup discovery, on the other, their marriage seems to offer a particularly fertile area of future research. Let us first imagine that a given prototypical point is specified in the Hilbert Space of the given kernel, which, using an applicable representer theorem [AMP09] can be achieved using an appropriate combination of the available data. Then, one can restrict the sensitivity of our method to be only with respect to a given type of differences. Let us now consider once more our example of traded stocks (see Appendix B.2); here, we could use as prototype a set of stocks that we know were affected to some degree only after the announcement of a successful vaccine. This would now allow us to find a description for those stocks that were affected the most in the same time and on the same way as the ones we chose as prototype. As yet another

application, one could even ask to describe the population to which a given prototype corresponds the most; now we could specify a set of stocks that were affected in a given way, and seek to describe the sub-population whose representation in Hilbert Space is the closest to that of the prototype.

On a different note, a related research direction would be to exploit the clustering-resembling formulation of our kernelised subgroup discovery method to enforce that the set of results is diverse, by additionally requiring the points in the Hilbert Space to be far not only from the dataset average, but also from any previously discovered subgroups. This would result with a collection of subgroups that each covers a different part of the data, and explains all exceptional sub-populations, with the added guarantee that each is substantially different than the other. We would ethus have a novel paradigm of named clustering, where each centroid corresponds to a describable sub-population, providing an significant improvement on unsupervised learning, when attribute information is available. Another approach for diverse results would be to explain the data in increasing detail, where on the first lever we consider the most consequential eigen-directions and we keep providing descriptions for increasingly less consequential ones. In this way, we would achieve a named variant of principal component analysis.

Our contributed algorithmic framework for the proposed optimisation of the related combinatorial problems combine together to form efficient methods which allow the practical application of subgroup discovery to a significantly broader domain. However, as is always the case, scaling combinatorial problems remains a constant challenge, and novel schemes can be of importance for tackling this issues. One first angle of attack would be to re-define what a tight optimistic estimator is, which is a rather loose definition, to one that utilised more information from each step of the optimisation, and disallows subgroups which would be impossible to form. In fact, although at first these advances seem to require a lot of up-keeping during the progression of our algorithm, the high complexity of combinatorial optimisation might justify some added costs.

Overall, in this thesis we provided a theoretical basis for subgroup discovery though a statistical perspective and adopted a multi-objective approach when we presented the results of each of the methods we developed. We have used our efficient and implementation of branch and bound, that has made practical to find the subgroup that optimises each adopted objective function. All our methods have been demonstrated on real-world data, whose structure ranges from either having an additional control variable, additional relations between them, or containing entities of virtually any possible structure. We therefore provide a generalised framework for the application of subgroup

discovery to novel kinds of datasets that were previously not feasible.

Appendix

A Delegated Proofs

In this appendix we provide proofs for our theoretical claims in this thesis.

A.1 Representative Subgroups

Lemma 3.2 (Optimal c-t path)**.** *Let* $\boldsymbol{e}_1 = (1,0)^T$ *and* $\boldsymbol{e}_2 = (0,1)^T$ *be the standard basis vectors of* $\mathbb{R}^2$*.*

i) Then the μ*-th element of the optimal c-t path is the class count of the first* μ *elements of* E*; formally,*

$$\boldsymbol{\pi}^{(\mu)} = \sum_{i=1}^{\mu} \boldsymbol{e}_{c_i} \qquad 0 < \mu \leq |Q| \qquad \textit{and} \qquad \boldsymbol{\pi}^{(\mu)} = \mathbf{0}\ . \tag{3.15}$$

ii) Moreover, the sequence ${f_{ct}}^{Q} \circ \boldsymbol{\pi}$*, with elements the* ${f_{ct}}^{Q}$ *values computed along the c-t path* $\boldsymbol{\pi}$*, is a concave sequence.*

Proof. Let $Q^*_\mu \subseteq Q$ be the set attaining the highest f_{ct} value among those with cardinality μ. We now reinterpret Eq. (3.14) as follows: the element $\boldsymbol{\pi}^{(\mu)}$ is equal to the index of the equi-count refinement set $\mathcal{R}_{\boldsymbol{\pi}^*}$ containing Q^*_μ. Within all sets with a fixed cardinality f_{c} remains constant, and Q^*_μ is the set with the maximal central tendency f_{t}; we can then show, similar to Proposition 3.1, that the maximiser of f_{t} contains the topmost μ target values. Altogether, $\boldsymbol{\pi}^{(\mu)}$ is exactly the class count of

$$Q^*_\mu \coloneqq \operatorname*{arg\,max}_{|R|=\mu,\, R\subseteq Q} f_{\mathrm{ct}}(R) = \operatorname*{arg\,max}_{|R|=\mu,\, R\subseteq Q} f_{\mathrm{t}}(R) = \bigcup_{i=1}^{\mu} \{\epsilon_i\}\ , \tag{A.1}$$

whose control class count is equal to the quantity in Eq. (3.15).

To show (ii) we proceed as follows. Since Q^*_μ contains the top-μ elements, we rewrite Eq. (3.10) as

$$(f_{\mathrm{ct}} \circ \boldsymbol{\pi})(\mu) = f_{\mathrm{ct}}(\boldsymbol{\pi}^{(\mu)}) = \alpha_{\mathrm{t}} \sum_{i=1}^{\mu} y_i - \alpha_{\mathrm{c}}\mu\ , \tag{A.2}$$

with discrete derivatives

$$\Delta_\mu(f_{\text{ct}} \circ \boldsymbol{\pi}) = f_{\text{ct}}(\boldsymbol{\pi}^{(\mu+1)}) - f_{\text{ct}}(\boldsymbol{\pi}^{(\mu)}) = \alpha_{\text{t}} y_{\mu+1} - \alpha_{\text{c}} \tag{A.3}$$

$$\Delta^2_\mu(f_{\text{ct}} \circ \boldsymbol{\pi}) = \alpha_{\text{t}}(y_{\mu+1} - y_\mu) \leq 0 \ , \tag{A.4}$$

where the last inequality holds because y_μ are decreasing. The negativity of the second discrete derivative, shows the concavity of the sequence. □

We next show Proposition A.3, which involves partitioning the CCS in compact regions surrounding the SST, within which the monotonicity of the factors ${f_{\text{ct}}}^Q$ and ${f_{\text{r}}}^Q$ remains constant, when computed along any horizontal or vertical sequences. All the sequences formed in this way increase toward the region boundary intersecting with the SST, and so no maximiser of ${f_{\text{cif}}}^Q$ can lie within these regions, except on the intersection of the region and the SST.

To show the above, we study both factors, starting with ${f_{\text{ct}}}^Q$.

Lemma A.1 (Domination of the ${f_{\text{ct}}}^Q$ factor)**.** *The impact value computed along any horizontal sequence is increasing until the abscissa π_1^* of the optimal c-t point, and decreasing afterwards. Similarly, the impact value computed along any vertical sequence is increasing until the ordinal π_2^* of the optimal c-t point, and decreasing afterwards. Formally,*

$${f_{ct}}^Q(\boldsymbol{I} + \boldsymbol{e}_k) \geq {f_{ct}}^Q(\boldsymbol{I}) \ , \qquad I_k < \pi_k^* \tag{A.5}$$

$${f_{ct}}^Q(\boldsymbol{I} + \boldsymbol{e}_k) \leq {f_{ct}}^Q(\boldsymbol{I}) \ , \qquad I_k \geq \pi_k^* \tag{A.6}$$

Proof. Denote the **optimal c-t path index** μ^* to be the index within the c-t path sequence attaining the maximum c-t value, $\boldsymbol{\pi}^{(\mu^*)} = \boldsymbol{\pi}^*$, so that

$$\begin{matrix} f_{\text{ct}}(\boldsymbol{\pi}^{(\mu+1)}) \geq f_{\text{ct}}(\boldsymbol{\pi}^{(\mu)}) & \mu < \mu^* \\ f_{\text{ct}}(\boldsymbol{\pi}^{(\mu+1)}) \leq f_{\text{ct}}(\boldsymbol{\pi}^{(\mu)}) & \mu \geq \mu^* \end{matrix} \ , \tag{A.7}$$

due to the concavity of the sequence $({f_{\text{cif}}}^Q \circ \boldsymbol{\pi})(\mu)$.

However, for any two consecutive points on the path, we can compute $f_{\text{ct}}(\boldsymbol{\pi}^{(\mu+1)}) - f_{\text{ct}}(\boldsymbol{\pi}^{(\mu)}) = \alpha_{\text{t}} y_{\mu+1} - \alpha_{\text{c}}$, which combined with Eq. (A.7) yields

$$\begin{matrix} \alpha_{\text{t}} y_{\mu+1} - \alpha_{\text{c}} \geq 0 & \mu < \mu^* \\ \alpha_{\text{t}} y_{\mu+1} - \alpha_{\text{c}} \leq 0 & \mu \geq \mu^* \end{matrix} \ . \tag{A.8}$$

Moreover, using Eq. (3.10) we can express the ${f_{\text{ct}}}^Q$ value of the point next

to $\boldsymbol{I}$ in CCS along dimension k as

$$f_{\mathrm{ct}}{}^{Q}(\boldsymbol{I}+\boldsymbol{e}_k) = \alpha_{\mathrm{t}} \sum_{k=1}^{2} \left(\sum_{\mathrm{i}=1}^{I_k} y_{\mathrm{i}}^{(k)} + y_{I_k+1}^{(k)} \right) - \alpha_{\mathrm{c}}(I_1+I_2)\ , \tag{A.9}$$

and so the difference between the $f_{\mathrm{ct}}{}^{Q}$ values of these neighbouring points becomes

$$f_{\mathrm{ct}}{}^{Q}(\boldsymbol{I}+\boldsymbol{e}_k) - f_{\mathrm{ct}}{}^{Q}(\boldsymbol{I}) = \alpha_{\mathrm{t}} y_{I_k+1}^{(k)} - \alpha_{\mathrm{c}}\ , \tag{A.10}$$

which is the quantity whose sign we study. According to Eq. (3.15), however, $\boldsymbol{\pi}$ is a sequence of single step increases $\boldsymbol{\pi}^{(\mu+1)} - \boldsymbol{\pi}^{(\mu)} = \boldsymbol{e}_{y_\mu^{(k)}}$, starting from the empty count $\mathbf{0}$. In other words, the μ-th element $\boldsymbol{\pi}^{(\mu)}$ of the sequence increases this class count that matches the class of the item in Q with the next greatest target value. This implies that at the optimal c-t path index μ^*, the optimal c-t path count $\boldsymbol{\pi}^* = \boldsymbol{\pi}^{(\mu^*)}$ per class amounts exactly to the number of items with the same control class and greater target value. Moreover, for each $I_k \geq \pi_k^*$ there exists a $\mu \geq \mu^*$ such that $y_\mu = y_{I_k+1}^{(k)}$, and similarly for each $I_k \leq \pi_k^*$ there exists a $\mu \leq \mu^*$ such that $y_\mu = y_{I_k+1}^{(k)}$. We can now combine the two equations Eq. (A.10), and Eq. (A.8), to show the claim of the lemma. □

We now proceed to show a similar behaviour of the f_{r} factor.

Lemma A.2 (Total Variation Domination)**.** *The composition of f_r with every horizontal sequence $\boldsymbol{h}_{\mathrm{i}}$, $\mathrm{i} = 0, \ldots, n_1(Q)$, and every vertical sequence $\boldsymbol{v}_{\mathrm{i}}$, $\mathrm{i}_1 = 0, \ldots, n_1(Q)$ forms the sequences $(f_r \circ \boldsymbol{h}_{\mathrm{i}})(\tau)$ and $(f_r \circ \boldsymbol{v}_{\mathrm{i}})(\tau)$; these are (i) uni-modal, (ii) attain a maximum at their intersection $(\mathrm{i},\mathrm{i})^T$ with the equi-representativeness ray $a(1,1)^T$, $a \geq 0$, and (iii) they are concave for the indices $\tau = 0, \ldots, \mathrm{i}$.*

Proof. We first focus on the horizontal sequences $(f_{\mathrm{r}} \circ \boldsymbol{h}_{\mathrm{i}})(\tau)$ for $\mathrm{i} = 0, \ldots, n_2(Q)$ and $\tau = 0, \ldots, m$. Notice that the $d_{\mathrm{TV}}(\boldsymbol{I})$ vanishes on the equi-representativeness ray $a(1,1)^T$, that is, when $I_1 = I_2$. Since the horizontal sequence $\boldsymbol{h}_{\mathrm{i}}$ has a fixed ordinal of i, the previous condition yields $\tau = \mathrm{i}$, which shows the correctness of (ii).

To prove the rest of this lemma, we study the continuous analogue of $(f_{\mathrm{r}} \circ \boldsymbol{h}_{\mathrm{i}})(\tau)$

$$\tilde{f}_{\mathrm{r}}(t) \coloneqq 1 - \left| \frac{1}{2} - \frac{t}{t+\mathrm{i}} \right|\ , \qquad\qquad t > 0\ , \tag{A.11}$$

which has first and second derivatives

$$\begin{aligned} \tilde{f}_{\mathrm{r}}{}'(t) &= \quad \operatorname{sign}\left(\frac{1}{2}-\frac{t}{t+\mathrm{i}}\right)\frac{\mathrm{i}}{(t+\mathrm{i})^2} \\ \tilde{f}_{\mathrm{r}}{}''(t) &= -2\operatorname{sign}\left(\frac{1}{2}-\frac{t}{t+\mathrm{i}}\right)\frac{\mathrm{i}}{(t+\mathrm{i})^3} \end{aligned} \qquad t \neq \mathrm{i}\,. \tag{A.12}$$

The sign of both quantities is controlled by the sign factor, which is negative when $t < \mathrm{i}$ and positive when $t > \mathrm{i}$, and so we can reach the conclusion that $\tilde{f}_{\mathrm{r}}{}'$ is increasing concave for $t < \mathrm{i}$ and decreasing convex for $t > \mathrm{i}$. Since $(f_{\mathrm{r}} \circ \boldsymbol{v}_{\mathrm{i}})(\tau) = \tilde{f}_{\mathrm{r}}(t)$, the above properties transfer to the discrete sequence $(f_{\mathrm{r}} \circ \boldsymbol{v}_{\mathrm{i}})$, as well. For vertical sequences, the symmetric argumentation can be used. □

We can now prove the sufficiency of SST, as claimed by Proposition A.3, by combining the two domination Lemmas A.1 and A.2.

Proposition A.3 (Sufficient Search Triangle)**.** *The maximum of the controlled impact function $f_{cif}{}^{Q}$ is attained at a point which lies in the (filled) triangle $\mathcal{T}(Q) \coloneqq \{(\pi_1^*, \pi_1^*), (\pi_2^*, \pi_2^*), \boldsymbol{\pi}^*\}$, with vertices the optimal c-t point $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*)$ and its horizontal and vertical projections onto the maximum $f_r{}^{Q}$ ray. We call this region the* ***sufficient search triangle****.*

Proof. For this proof we show that every point outside the SST is dominated by one within $\mathcal{T}(Q)$. For this we distinguish two cases, depending on whether the c-t optimal point is above or below the maximum representativeness line.

Assume the optimal c-t point lies above the maximum $f_{\mathrm{r}}{}^{Q}$ ray. The point μ^*, along with the maximum representativeness ray, partition the CCS in the 6 regions shown in Fig. A.1, each of which has a non empty intersection with $\mathcal{T}(Q)$. We now show that the maxima of f_{cif} over all the points in these regions, lie on this intersection, and therefore also in the SST.

We first study A_{SW}: the points on the diagonal maximise $f_{\mathrm{r}}{}^{Q}$, while at the same time $f_{\mathrm{ct}}{}^{Q}$ is dominated by the SST point (π_1^*, π_1^*), therefore maximising f_{cif} altogether. Similarly within A_{NE}, we can show that f_{cif} is maximised by $(\pi_2^*, \pi_2^*) \in \mathcal{T}(Q)$.

Within regions A_{W} and A_{N}, both terms increase along each west-to-east and north-to-south path, respectively; these paths lead to a point of $\mathcal{T}(Q)$ that dominates all the rest on the traversed path. Within A_{SW}, each west-to-east path ends up in a point of A_{N}, which is itself dominated by a point of SST.

Finally, every south-to-north path of A_{SE} leads to either a point of $\mathcal{T}(Q)$ directly, or to one in the dominated A_{NE}. We thus showed that no point of $\mathcal{I}(Q) \setminus \mathcal{T}(Q)$ can maximise f_{cif}.

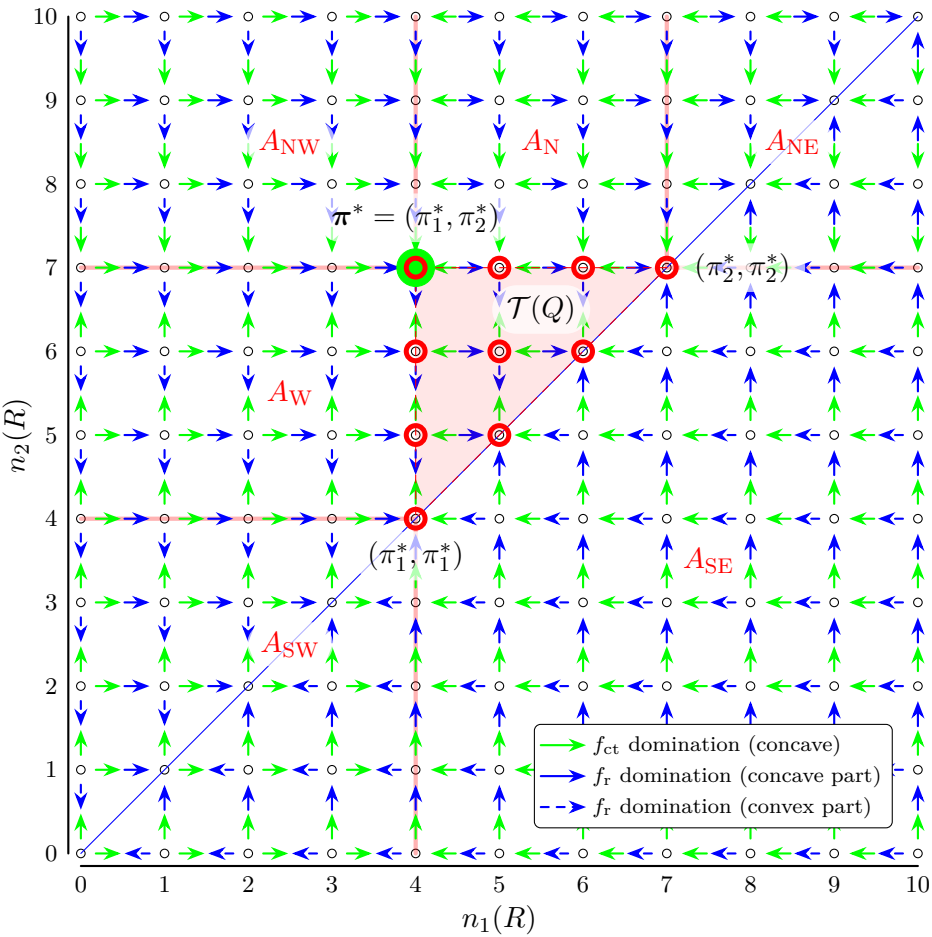


Figure A.1: Domination relations for a c-t optimal point $\boldsymbol{\pi}^*$ above the maximum ${f_{\mathrm{r}}}^{Q}$ ray: the arrows point to the greater factor value. The $\mathcal{T}(Q)$ partitions $\mathcal{I}$ in the 6 marked areas.

We work likewise if $\boldsymbol{\pi}^*$ lies below the maximum ${f_{\mathrm{r}}}^{Q}$ ray. □

Proposition A.4 (Concavity of ${f_{\mathrm{cif}}}^{Q}$ along sequence)**.** *Consider the values of the controlled impact function ${f_{cif}}^{Q}$ as they are computed along a horizontal sequence $\boldsymbol{h}_{\mathrm{i}_2}$; these form the sequence $({f_{cif}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})(\mu)$, which for $\mu \leq \mathrm{i}_2$ is a concave sequence preceding the maximum ${f_r}^{Q}$ ray. Similarly, $({f_{cif}}^{Q} \circ \boldsymbol{v}_{\mathrm{i}_1})(\mu)$ is a concave sequence for $\mu \leq \mathrm{i}_1$.*

Proof. To prove this statement we employ the concavity of the sequences formed as the two factors ${f_{\mathrm{ct}}}^{Q}$ and ${f_{\mathrm{r}}}^{Q}$ are computed along the horizontal and vertical sequences. We first treat the horizontal sequences, along which the entire ${f_{\mathrm{ct}}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2}$ is concave, and so is $({f_{\mathrm{r}}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})(\mu)$ for the indices $\mu = 0, \ldots, \mathrm{i}_2$, according to Lemmas 3.2 and A.1, respectively.

Additionally, all factors are positive (or can be made so by adding an appropriate constant term) and so, raising the elements of the sequences to a power in $(0, 1]$ preserves concavity. Multiplying the two re-weighted sequences yields

$$\left(({f_{\mathrm{ct}}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})^{\gamma}({f_{\mathrm{r}}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})^{1-\gamma})\right)(\mu) = ({f_{\mathrm{cif}}}^{Q} \circ \boldsymbol{h}_{\mathrm{i}_2})(\mu)\ , \tag{A.13}$$

which is concave as the multiplication of two concave, positive sequences, therefore showing the concavity of the sequence of impact function values computed along the specified horizontal sub-sequence. Similarly we can work for vertical sequences.

Note that in our analysis we seamlessly interchange continuous and discrete convexity definitions. This is enabled by the uni-variate nature of the functions involved, since their discrete counterparts corresponds to sampling on regular intervals. Indeed, on one hand it can be shown that regular sampling of a uni-variate convex function yields a convex sequence [Yüc02]. As a sufficiently applicable inverse for our needs, we can show that for any convex sequence, there exists at least one convex function with the same values at the sampled points and continuous second derivative; one such function results from cubic spline interpolation fitted on the sequence values. □

A.2 Robustly Connected Subgroups

Proposition 4.2 (Piece-wise Linear Estimate)**.** *For the piece-wise linear function of Eq.* (4.13)

i) $\kappa_U^i \leq \hat{\kappa}_U^i$, *for all* $0 \leq i \leq |U|$

ii) $\kappa_U^i = \hat{\kappa}_U^i$, *for* $i \in \{0, n_0, \ldots, n_{K_U}\}$

Proof. We first prove the first part of the proposition. Recall that the vertices of U are indexed in decreasing order of their coreness in the induced graph, so that $\kappa_U(v_i) \geq \kappa_U(v_{i+1})$ for all $1 \leq i < |U|$. Let $U_i \coloneqq \{v_1, \ldots, v_i\}$ be the subset of U containing the vertices with the top-i core numbers in $G[U]$. Then, we can write $\hat{\kappa}_U^i \equiv \kappa_U(U_i)$, where we extend the definition of Eq. (4.4) to accept the vertex set T over which the sum runs

$$\kappa_U(T) \coloneqq \sum_{u \in T} \kappa_U(v) = |T| \bar{\kappa}_U(T) \; . \tag{A.14}$$

We can now show that the coreness of U_i is the highest among all subsets $T \subseteq U$ with fixed cardinality $|T| = i$. In particular, for all $T \neq U_i$ we can always build a sequence $T^{(0)}, \ldots, T^{(\tau)}$ of subsets of U that starts with $T^{(0)} = T$ and ends with $T^{(\tau)} = U_i$, whose elements have increasing core sums $\kappa_U(T^{(\tau)}) < \kappa_U(T^{(\tau+1)})$. Such a sequence can be constructed by repeatedly swapping two vertices $u \in T^{(\tau)} \setminus U_i$ and $v \in U_i \setminus T^{(\tau)}$ with $\kappa_U(u) < \kappa_U(v)$, so that $T^{(\tau+1)} = (T^{(\tau)} \setminus \{v\}) \cup \{u\}$. Then for each $0 \leq \tau < \tau$ we have

$$\kappa_U\big(T^{(\tau+1)}\big) = \kappa_U\big(T^{(\tau)}\big) + \kappa_U(u) - \kappa_U(v) > \kappa_U\big(T^{(\tau)}\big) \; , \tag{A.15}$$

and summing the above relations for each index τ results in

$$\hat{\kappa}_U^i \equiv \kappa_U(U_i) \geq \kappa_U(T) \; , \qquad \text{for all } T \subseteq U \text{ with } |T| = i \, . \tag{A.16}$$

Let T_i^* be the core sum maximiser over subsets $T \subseteq U$ with $|T| = i$:

$$T_i^* \coloneqq \underset{T \subseteq U, \, |T| = i}{\arg\max} \; \kappa_T(T) \; , \qquad \text{so that } \kappa_{T_i^*}(T_i^*) =: \kappa_U^i \, . \tag{A.17}$$

The last two Eqs. (A.16) and (A.17) can now be combined as

$$\hat{\kappa}_U^i = \kappa_U(U_i) \geq \kappa_U(T_i^*) \geq \kappa_{T_i^*}(T_i^*) =: \kappa_U^i \, , \tag{A.18}$$

where the second inequality follows by Lemma 4.1. This concludes the proof of part i).

To prove part ii), we first consider all vertex subsets $T \subseteq U$ with cardinality $|T| = n_k$, equal to the size of some k-core of $G[U]$. Let T_k be one of those subsets for which $G[T_k]$ achieves the highest total coreness $\kappa_U^{n_k} = \kappa_{T_k}$. We need to show that this value is equal to the bound of Eq. (4.13): $\kappa_U^{n_k} = \kappa_{T_k} = \hat{\kappa}_U^{n_k}$.

This optimal subset is exactly $T_k \equiv E_U(k)$: the vertices in the k-core of

$G[U]$. Now κ_{T_k} becomes the total coreness of

$$G[T_k] = G[U][T_k] = G[U][E_U(k)] = H_U(k) \ , \tag{A.19}$$

where the first equality follows from basic graph properties and the latter from the definition of the k-core of $G[U]$, which means that for every vertex $u \in T_k$ it holds that $\kappa_{T_k}(u) \geq k$.

However, due to the hierarchical structure of the k-cores, we can write a similar inequality for each nested core

$$\kappa_U(v) \geq \kappa_{T_\lambda}(v) \geq \lambda \ , \qquad \text{for all } v \in T_\lambda \, , \, k \leq \lambda \leq K_U \, , \tag{A.20}$$

where $T_\lambda \coloneqq E_U(\lambda)$ and the first inequality follows from Lemma 4.1 since $T_\lambda \subseteq U$. For each vertex v we can select the strictest inequality described in Eq. (A.20) and then sum them up. Out of those, there are $|E_{k+1}(T)\,| - |E_k(T)\,| = |E_{k+1}(U)\,| - |E_k(U)\,| =: n_{k+1} - n_k$ many with a bound of λ, for each $k \leq \lambda K_U$, which add up to $\kappa_U^{n_k} \geq \kappa_{T_k}(T_k) \geq \sum_{\lambda=k}^{K_U+1}(n_{\lambda+1} - n_\lambda)\lambda$. We can now combine this inequality with the proven claim in part i) and the quantity in Eq. (4.13) to show the required equality. □

Corollary 4.3 (Optimistic Estimate)**.** *The quantity $\hat{\phi}_U^*(\gamma)$ is an optimistic estimator of $f_{di}(U;\gamma)$. In addition, $\hat{\phi}_U^*$ becomes tight for $\gamma = 1/2$.*

$$\hat{\phi}_U^*(\gamma) \coloneqq \max_{0<i\leq|U|} \left(\frac{i}{|V|}\right)^{1-\gamma} \left(\frac{\hat{\kappa}_U^i}{i} - \bar{\kappa}_V\right)^{\gamma} . \tag{4.16}$$

Proof. To show the first part we need to prove that for a given subset U and for all its subsets $T \subseteq U$ it is $\hat{\phi}_U^*(\gamma) \geq f_{\text{di}}(T;\gamma)$. We study an arbitrary $T \subseteq U$ and notice that

$$\hat{\kappa}_U^{|T|} \geq \kappa_U^{|T|} \geq \kappa_T(T) \ , \tag{A.21}$$

where the first inequality is due to part i) of Proposition 4.2 and the latter stems from the definition of $\kappa_U^{|T|}$ in Eq. (4.11). We can now transform the first and last sides of Eq. (A.21) with monotonicity preserving operations: we first subtract the constant $\kappa_V(V)/|V|$, then raise to the positive power of

γ and multiply with the positive quantity $(|T|/|V|)^{1-\gamma}$; we thus derive

$$\hat{\phi}_U(|T|;\gamma) \coloneqq \left(\frac{|T|}{|V|}\right)^{1-\gamma} \left(\frac{\hat{\kappa}_U^{|T|}}{|T|} - \bar{\kappa}_V(V)\right)^{\gamma} \tag{A.22}$$

$$\geq \left(\frac{|T|}{|V|}\right)^{1-\gamma} \left(\bar{\kappa}_T - \bar{\kappa}_V(V)\right)^{\gamma} =: f_{\text{di}}(T;\gamma)\,. \tag{A.23}$$

By the definition of Eq. (4.15) it is also $\hat{\phi}_U^*(\gamma) \geq \hat{\phi}_U(T;\gamma)$, which combined with the previous result proves the first half of the claim.

For the second part we need to prove that any subset U has a sub-subset $T \subseteq U$ such that $\hat{\phi}_U^* = f_{\text{di}}(T)$; in fact such an T corresponds to adding a complete core $T = E_k(U)$. To show this we first rewrite our optimistic estimator, which in the case of $\gamma = 1/2$ becomes

$$\hat{\phi}_U(i) \coloneqq \alpha\hat{\kappa}_U^i - \beta i\,, \qquad \text{with} \quad \begin{array}{l} \alpha \coloneqq 1/|V| > 0 \\ \beta \coloneqq \dfrac{\bar{\kappa}_V}{|V|} \geq 0 \end{array}\,. \tag{A.24}$$

Taking the first order finite difference of Eq. (A.24) we get

$$\begin{array}{rl} \Delta\hat{\phi}_U(i) \coloneqq \hat{\phi}_U(i+1) - \hat{\phi}_U(i) & n_{\kappa+1} \leq i < n_\kappa \\ = \alpha\kappa - \beta & 0 \leq \kappa \leq K_U \end{array}\,. \tag{A.25}$$

Additionally, the core numbers κ decrease as i increases, with changes happening if and only if the segment changes. Therefore, if the finite difference becomes negative, this must happen either at a segment boundary, or during an entire segment; in any case, there is an index $i^* = n_{\kappa^*}$ for which the sequence $\hat{\phi}_U(i^*) = \hat{\phi}_U^*$. This point corresponds to $\hat{\kappa}_U^{n_{\kappa^*}}$ which is equal to $\kappa_U^{n_{\kappa^*}}$ according to Proposition 4.2 (part ii)), so that then $\hat{\phi}_U^* = \hat{\phi}_U(i^*) = f_{\text{di}}(T)$ for $T = E_{\kappa^*}(U)$. □

A.3 Kernelised Subgroup Discovery

Lemma 5.1. *Let $m_Q \coloneqq |Q|$ be the cardinality of any entity subset. Then we can write our objective of Eq.* (5.7) *as*

$$J(Q;\kappa,\gamma) = a_t^{\gamma-2}(m_Q)\mathbf{z}_Q{}^\top K \mathbf{z}_Q\,, \tag{5.9}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ *is the Gramian* $\mathbf{K}_{i,j} \coloneqq \kappa(\epsilon_i, \epsilon_j)$ *and*

$$\mathbf{z}_Q \coloneqq \mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}\,, \tag{5.10}$$

for $\mathbf{e} \coloneqq (1, \ldots, 1) \in \mathbb{R}^n$ *the vector of all ones and* $\mathbf{x}_Q \coloneqq (\mathbb{1}[\epsilon_i \in Q])_{i=1}^n$ *the characteristic vector of set* Q*; here we denote* $\mathbb{1}[\cdot]$ *the characteristic*[1] *function.*

Proof. Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the Gramian of the entries of E with respect to the kernel κ under some arbitrary but fixed ordering, and denote $\mathbf{x}_{Q'_t}$ the characteristic vector of Q'_t. We can now revisit the empirical estimator of Eq. (5.6) and write

$$\frac{1}{|Q||Q'_t|} \sum_{\substack{\epsilon \in Q, \\ \epsilon' \in Q'_t}} \kappa(\epsilon, \epsilon') = \frac{1}{|Q||Q'_t|} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}[\epsilon_i \in Q]\mathbb{1}[\epsilon_j \in Q'_t]\mathbf{K}_{i,j} = \frac{{\mathbf{x}_Q}^\top \mathbf{K} \mathbf{x}_{Q'_t}}{m_Q \cdot m_{Q'_t}}\,, \tag{A.26}$$

where $m_{Q'_t} \coloneqq |Q'_t|$. Using a similar process for the other two terms of Eq. (5.6) we may rewrite it as

$$\widehat{\mathrm{MMD}}^2(Q, Q'_t) = \left(\frac{1}{m_Q}\mathbf{x}_Q - \frac{1}{m_{Q'_t}}\mathbf{x}_{Q'_t}\right)^\top \mathbf{K} \left(\frac{1}{m_Q}\mathbf{x}_Q - \frac{1}{m_{Q'_t}}\mathbf{x}_{Q'_t}\right). \tag{A.27}$$

We now observe that $\mathbf{x}_{Q'_{\mathrm{ano}}} = \mathbf{e}$ and $\mathbf{x}_{Q'_{\mathrm{con}}} = \mathbf{e} - \mathbf{x}_Q$, and therefore $m_{Q'_{\mathrm{ano}}} = n$ and $m_{Q'_{\mathrm{con}}} = n - m_Q$, and therefore

$$\frac{1}{m_Q}\mathbf{x}_Q - \frac{1}{m_{Q'_{\mathrm{ano}}}}\mathbf{x}_{Q'_{\mathrm{ano}}} = \frac{1}{m_Q}\left(\mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}\right) = \frac{\mathbf{z}_Q}{a_{\mathrm{ano}}(m_Q)} \tag{A.28}$$

[1] That is, $\mathbb{1}[\cdot] = 1$ if the condition $\cdot$ is satisfied and 0 otherwise.

and similarly

$$\frac{1}{m_Q}\mathbf{x}_Q - \frac{1}{m_{Q'_{\text{con}}}}\mathbf{x}_{Q'_{\text{con}}} = \frac{\mathbf{x}_Q}{m_Q} - \frac{\mathbf{e}-\mathbf{x}_Q}{n-m_Q} = \frac{\mathbf{x}_Q(n-m_Q)-m_Q(\mathbf{e}-\mathbf{x}_Q)}{m_Q(n-m_Q)} = \frac{n}{m_Q(n-m_Q)}\left(\mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}\right) = \frac{\mathbf{z}_Q}{a_{\text{con}}(m_Q)}\,. \quad \text{(A.29)}$$

We can now replace Eqs. (A.28) and (A.29) into Eq. (A.27) to get

$$\widehat{\text{MMD}}^2(Q,Q'_t) = a_t^{-2}(m_Q)\cdot \mathbf{z}_Q{}^\top \mathbf{K}\mathbf{z}_Q \quad \text{(A.30)}$$

and finally multiply both sides of Eq. (A.30) with $a_t^{\gamma}(m_Q)$ to show the quantity in the claim. □

We next prove the claim in the bound of Lemma 5.2. We first observe that

$$\|\mathbf{z}_Q\|_2^2 = \left\|\mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}\right\|_2^2 = \underbrace{\|\mathbf{x}_Q\|_2^2}_{m_Q} - 2\frac{m_Q}{n}\underbrace{\mathbf{e}^\top\mathbf{x}_Q}_{m_Q} + \left(\frac{m_Q}{n}\right)^2\underbrace{\|\mathbf{e}\|_2^2}_{n} = \frac{m_Q(n-m_Q)}{n} = a_{\text{con}}(m_Q)\,, \quad \text{(A.31)}$$

where we used that $\mathbf{x}_Q{}^\top\mathbf{x}_Q = \mathbf{e}^\top\mathbf{x}_Q = m_Q$. Since $\mathbf{K}$ is symmetric positive (semi-)definite, it is also diagonalisable with non-negative eigenvalues and we can always compute its spectral decomposition $\mathbf{S}\mathbf{\Lambda}\mathbf{S}^\top = \mathbf{K}$, where $\mathbf{S}\in\mathbb{R}^{m\times k}$ is orthonormal with $\mathbf{S}^\top\mathbf{S} = \mathbf{I}$ and $\mathbf{\Lambda}\in\mathbb{R}^{k\times k}$ is diagonal with $\mathbf{\Lambda}_{i,i} = \lambda_i$ and $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$. Using the result of Eq. (A.31) we can now write the quadratic form of Eq. (5.9) as

$$\mathbf{z}_Q{}^\top\mathbf{K}\mathbf{z}_Q = a_{\text{con}}(m_Q)\frac{\mathbf{z}_Q{}^\top\mathbf{K}\mathbf{z}_Q}{\mathbf{z}_Q{}^\top\mathbf{z}_Q} = a_{\text{con}}(m_Q)\sum_{i=1}^{k}\lambda_i\cos^2\phi_i(Q)\,, \quad \text{(A.32)}$$

where $\phi_i(Q) \coloneqq \measuredangle(\mathbf{z}_Q,\mathbf{v}_i)$ is the angle between the vector $\mathbf{z}_Q$ and the i-th

eigenvector. Before we proceed with the actual proof, we will need the following lemma.

Lemma A.5. *The cosine of the angle between the vector* $\mathbf{z}_Q$ *and any eigenvector* $\mathbf{v}_i$ *is bounded by*

$$u_i \coloneqq \max_{\substack{R\subseteq Q\\|R|=m}} \cos^2\phi_i(R) = \frac{1}{a_{con}(m)}\left(\max\left\{\mathbf{e}_{:m}{}^\top\mathbf{v}_{i\uparrow[Q]}, \mathbf{e}_{:m}{}^\top\mathbf{v}_{i\downarrow[Q]}\right\} - \frac{m}{n}\mathbf{e}^\top\mathbf{v}_i\right)^2. \quad \text{(A.33)}$$

Proof. We first study the maximum without the square, for which we get

$$\max_{\substack{R\subseteq Q\\|R|=m}} \cos\phi_i(R) = \max_{\substack{R\subseteq Q\\|R|=m}} \frac{\mathbf{v}_i{}^\top\mathbf{z}_R}{\|\mathbf{z}_R\|_2} = \frac{1}{\sqrt{a_{\mathrm{con}}(m)}}\left(-\frac{m}{n}\mathbf{e}^\top\mathbf{v}_i + \max_{\substack{R\subseteq Q\\|R|=m}} \mathbf{x}_R{}^\top\mathbf{v}_i\right) = \frac{\mathbf{e}_{:m}{}^\top\mathbf{v}_{i\uparrow[Q]} - \frac{m}{n}\mathbf{e}^\top\mathbf{v}_i}{\sqrt{a_{\mathrm{con}}(m)}}, \quad \text{(A.34)}$$

where we used Eq. (A.31) to normalise $\mathbf{z}_R$ under the fixed cardinality condition $|R| = m$, and the fact that maximising the inner product of $\mathbf{v}_i$ and the characteristic function $\mathbf{x}_R$ of any subset of $R \subseteq Q$ amounts to summing the m greatest elements of $\mathbf{v}_i$. This is easy to show e.g., by contradiction using induction, in which one could start with any other candidate characteristic vector $\mathbf{x}$ and repeatedly swap each of its elements with another that corresponds to an element in $\mathbf{v}_i$ with greater value.

In a similar fashion we can also show that the minimum value of the un-squared cosine is attained at

$$\min_{\substack{R\subseteq Q\\|R|=m}} \cos\phi_i(R) = \frac{\mathbf{e}_{:m}{}^\top\mathbf{v}_{i\downarrow[Q]} - \frac{m}{n}\mathbf{e}^\top\mathbf{v}_i}{\sqrt{a_{\mathrm{con}}(m)}}. \quad \text{(A.35)}$$

Using the fact that for any function f it holds that

$$\max_x f^2(x) = \left[\max\{\max_x f(x), \min_x f(x)\}\right]^2, \quad \text{(A.36)}$$

we can now derive the maximum of Eq. (A.33) as the squared maximum

between the quantities of Eqs. (A.34) and (A.35). □

Summarising Lemma A.5, it provides bounds $u_i \geq \max_{R\subseteq Q,|R|=m} \cos^2 \phi_i(R)$ for each $i = 1, \ldots, k$, where k is the rank of $\mathbf{K}$. We can now combine these bounds as follows.

Lemma 5.2. *Given any integer constant $\rho < k$, an upper bound for the problem in Eq.* (5.12) *is*

$$\hat{f}_t(Q;\kappa,\gamma,m) = a_t^{\gamma-2}(m_Q)a_{con}(m_Q)\left(\sum_{i=1}^{\rho}\lambda_i \min\{u_i,\vec{u}_i\} + \lambda_{\rho+1}\vec{u}_{\rho+1}\right), \quad (5.13)$$

where $\vec{u}_i \coloneqq \max\left\{0, 1-\sum_{j=1}^{i-1} u_j\right\}$ *and*

$$u_i \coloneqq \frac{\left(\max\left\{\mathbf{e}_{:m}{}^\top \mathbf{v}_{i\uparrow[Q]}, \mathbf{e}_{:m}{}^\top \mathbf{v}_{i\downarrow[Q]}\right\} - \frac{m}{n}\mathbf{e}^\top\mathbf{v}_i\right)^2}{a_{con}(m)}. \quad (5.14)$$

Proof. Let R be an arbitrary subset of Q with fixed cardinality $|R| = m$. We now focus in the formulation of Eq. (5.13) and consider only the factor in the sum. We further observe that for some $\rho' = \min\{r \mid \vec{u}_r < u_r\} \cup \{\rho\}$

$$\sum_{i=1}^{\rho}\lambda_i \min\{u_i,\vec{u}_i\} + \lambda_{\rho+1}\vec{u}_{\rho+1} = \sum_{i=1}^{\rho'}\lambda_i u_i + \lambda_{\rho'+1}\vec{u}_{\rho'+1}, \quad (A.37)$$

That is, the index $\rho' \leq \rho$ is the first index for which the coefficients u_i used so far would sum up to a number greater than 1, and thus ensures that $\sum_{i=1}^{n} u_i + \vec{u}_{\rho'} = 1$. Notice that this condition would also be satisfied if we kept adding more terms than ρ', but then the coefficients $\rho > \rho' + 1$ would vanish due to the minimum in the formulation of Eq. (5.13).

We now assume that the full basis of the eigen-space $\mathbf{S} \in \mathbb{R}^{n\times n}$ is available, and therefore the angles $\phi_i(R)$ of Eq. (A.32) are well defined for $i = 1, \ldots, n$. Importantly, since the cosines of these angles are directional cosines of a vector and each vector of an orthonormal basis, it must be that their squares sum to 1. An easy way to show this is using the orthonormality $\mathbf{S}^\top\mathbf{S} = \mathbf{I}$ in

the norm of a unit vector

$$\frac{\mathbf{z}_R}{\|\mathbf{z}_{R\|_2}}^\top \frac{\mathbf{z}_R}{\|\mathbf{z}_{R\|_2}} = 1 \iff$$

$$\frac{\mathbf{z}_R}{\|\mathbf{z}_{R\|_2}}^\top \mathbf{S}\mathbf{S}^\top \frac{\mathbf{z}_R}{\|\mathbf{z}_{R\|_2}} = 1 \iff$$

$$\left\| \mathbf{S}^\top \frac{\mathbf{z}_R}{\|\mathbf{z}_{R\|_2}} \right\|_2^2 = 1 \iff \left\| \Big(\cos \phi_i(R) \Big)_{i=1}^n \right\|_2^2 = 1 \iff$$

$$\sum_{i=1}^{n} \cos^2 \phi_i(R) = 1\,, \quad \text{(A.38)}$$

where we used the notation $(\cdot)_{i=1}^n$ to define a vector based on its elements.

We now define the remaining squared directional cosines of $\mathbf{z}_R$ in the eigenspace of $\mathbf{K}$ as

$$\eta_i(R) \coloneqq \sum_{j=i}^{\rho} \cos^2 \phi_j(R) = 1 - \sum_{j=1}^{i-1} \cos^2 \phi_j(R)\,. \quad \text{(A.39)}$$

We can now write for the sum in the objective formulation of Eq. (A.32) for R

$$\sum_{i=1}^{k} \lambda_i \cos^2 \phi_i(R) \leq$$

$$\sum_{i=1}^{\rho'} \lambda_i \cos^2 \phi_i(R) + \lambda_{\rho'+1} \sum_{i=\rho'+1}^{k} \cos \phi_i(R) \leq$$

$$\sum_{i=1}^{\rho'} \lambda_i \cos^2 \phi_i(R) + \lambda_{\rho'+1} \eta_{\rho'+1}(R)\,, \quad \text{(A.40)}$$

It now suffices to bound the right hand side of Eq. (A.40) with the right hand side of Eq. (A.37):

$$\sum_{i=1}^{\rho'} \lambda_i u_i + \lambda_{\rho'+1} \vec{u}_{\rho'+1} \geq$$

$$\sum_{i=1}^{\rho'} \lambda_i \cos^2 \phi_i(R) + \lambda_{\rho'+1} \eta_{\rho'+1}(R) \quad \text{(A.41)}$$

$$\iff \quad \text{(A.42)}$$

$$\sum_{i=1}^{\rho'} \lambda_i \left(u_i - \cos^2 \phi_i(R)\right) \geq$$

$$\lambda_{\rho'+1} \left(\eta_{\rho'+1}(R) - \vec{u}_{\rho'+1}\right) \geq$$

$$\lambda_{\rho'+1} \left(\cancel{1} - \sum_{i=1}^{\rho'} \cos^2 \phi_i(R) - \cancel{1} + \sum_{i=1}^{\rho'} u_i\right) \geq$$

$$\lambda_{\rho'+1} \sum_{i=1}^{\rho'} \left(u_i - \cos^2 \phi_i(R)\right) , \quad \text{(A.43)}$$

where the second inequality always holds, because $u_i \geq \cos^2 \phi_i(R)$ for all i, due to Lemma A.5, and therefore the difference in each term remains positive; since also $\lambda_i \geq \lambda_{\rho'+1}$, for all $i = 1, \ldots, \rho'$, the right hand side of Eq. (A.41) is not less than its left hand side.

By proving the inequality of Eq. (A.41) we showed that for any $R \subseteq Q$ with $|R| = m$

$$\sum_{i=1}^{k} \cos^2 \phi_i(R) \leq \sum_{i=1}^{\rho} \lambda_i \min\{u_i, \vec{u}_i\} + \lambda_{\rho+1} \vec{u}_{\rho+1} \overset{\cdot a_t^{\gamma-2} a_{\text{con}}}{\Longrightarrow} \quad \text{(A.44)}$$

$$a_t^{\gamma-2}(R) \mathbf{z}_R{}^\top \mathbf{K} \mathbf{z}_R \leq \hat{f}_t(Q; \kappa, m) \iff \quad \text{(A.45)}$$

$$J(R; \kappa, \gamma) \leq \hat{f}_t(Q; \kappa, \gamma, m) . \quad \text{(A.46)}$$

We can now show that $\hat{f}_t(Q; \kappa, \gamma, m)$ is an admissible bound for the fixed–cardinality objective of Eq. (5.12) by taking the maximum over all R with cardinality m. □

A.4 A Structure-Aware Graph Kernel

Lemma 6.1 (Matrix-Vector Operation)**.** *The matrix-vector operation of the edge similarity matrix of Eq.* (6.14) *can be computed in* $O(n'^2 n'' + n' n''^2)$ *as*

$$\mathbf{W}_\times \mathbf{x} = \text{vec}\Big[\mathbf{T} \circ \big(\mathbf{A}''(\mathbf{T} \circ \text{mat}[\mathbf{x}]\,)\mathbf{A}''^\top\big)\Big] , \quad \text{(6.15)}$$

where $\mathbf{T} \coloneqq \text{mat}[\mathbf{k}]$ *is the matricisation of Eq.* (6.13)*.*

Proof. We start with Eq. (6.14) and observe that

$$\begin{aligned}\mathbf{W}\mathbf{x} &= \left((\mathbf{A}' \otimes \mathbf{A}'') \circ (\mathbf{k}\mathbf{k}^\top)\right)\mathbf{x} && \text{(A.47)}\\ &= \operatorname{diag}(\mathbf{k})(\mathbf{A}' \otimes \mathbf{A}'')(\mathbf{x} \circ \mathbf{k}) && \text{(A.48)}\\ &= \operatorname{vec}[\mathbf{k}] \circ \operatorname{vec}\left[\operatorname{mat}[\mathbf{k}]\mathbf{A}''\operatorname{mat}[\mathbf{k} \circ \mathbf{x}]\mathbf{A}'^\top\right] && \text{(A.49)}\\ &= \operatorname{vec}\left[\operatorname{mat}[\mathbf{k}] \circ (\mathbf{A}''(\operatorname{mat}[\mathbf{k}] \circ \operatorname{mat}[\mathbf{x}])\mathbf{A}')\right] \ , && \text{(A.50)}\end{aligned}$$

where in the second line we used Eq. (6.12), in the third the identity of Eq. (6.6), and in the last we used basic properties of the vectorisation/matricisation operations.

This operation only involves Hadamard products and matrix-matrix products for matrices of size $n' \times n''$; assuming dense matrices, as a worst case scenario, we can derive the required complexity. □

Lemma 6.3 (Operation Complexity)**.** *The matrix-vector operation* $\mathbf{W}_\times\mathbf{x}$ *for the SUSAN kernel whose vertex kernel has a bounded support with bandwidth* δ *can be computed in exactly*

$$c = (2n'+2n''+1)\sum_{i=0}^{K'} b'_i \sum_{j=\max(0,\lambda-\delta)}^{\min(K'',i+\delta)} b''_j - n'n'' \tag{6.19}$$

floating point operations, which is $O\big((\delta+1)(n'+n'')B^2\big)$*, for* B *the maximal size among all* b'_k*,* b''_k*. In contrast, a naïve computation requires* $2n'n''(n'+n'')$ *such operations.*

Proof. Using Lemma 6.1, we can compute the matrix vector operation as per Eq. (6.15). This involves two Hadamard products and two matrix multiplications on symmetric matrices, and therefore the order of computation does not matter.

The computation involves the two block-ordered adjacency matrices $\mathbf{A}'$, $\mathbf{A}''$, with row- and column-block sizes equal to $\mathbf{b}'$ and $\mathbf{b}''$, respectively.

The innermost Hadamard product costs one computation per element for each non-zero block of $\mathbf{T}$, which equals

$$\sum_{i=0}^{K''} b''_i \sum_{j=\max(0,i-\delta}^{\min(K',i+\delta)} b'_j \ . \tag{A.51}$$

The result is then multiplied from left with matrix $\mathbf{A}''$; the resulting matrix has $K''+1 \times K'+1$ blocks, each entry of which requires operations equal to an inner product with size equal to the non-zero elements in $\mathbf{X}$. This

amounts to

$$\sum_{\lambda=0}^{K''} b''_\lambda \sum_{i=0}^{K'} b'_i \left(2 \sum_{j=\max(0,\lambda-\delta)}^{\min(K'',i+\delta)} b''_j - 1 \right) = \tag{A.52}$$

$$2n'' \sum_{i=0}^{K'} b'_i \sum_{j=\max(0,\lambda-\delta)}^{\min(K'',i+\delta)} b''_j - n'n'' \,, \tag{A.53}$$

where in the first line we use the fact that the inner product of two n-sized matrices costs $2n-1$ floating point operations, and in the second that the sum $\sum_{j=0}^{K'} b'_i = n'$, and similarly for n''.

The resulting matrix is then multiplied with $\mathbf{A}'$, but only the nonzero elements of the result are taken into account. For each of the non-zero entries an n'-sized inner product is required, which amounts to exactly

$$(2n'-1) \sum_{i=0}^{K'} b'_i \sum_{j=\max(0,\lambda-\delta)}^{\min(K'',i+\delta)} b''_j \tag{A.54}$$

floating point operations. Finally, a second Hadamard product ensues. The total number of computations follows from the addition of the above quantities.

The total number of operations for the naïve implementation follows from the addition of the cost of two Hadamard products to that of two matrix-matrix multiplications, adding up to

$$2(\underbrace{n'n''}_{\circ}) + \underbrace{n'(2n'-1)n''}_{\text{first } \cdot} + \underbrace{n'(2n''-1)n''}_{\text{second } \cdot} \,. \tag{A.55}$$

□

B Experiment Details and Used Datasets

B.1 Robustly Connected Subgroups

In this section we provide more details in the approach we used to create each dataset.

The datasets `Facebook`, `Twitter` and `Google+` were created by aggregating the ego-networks accessible from the SNAP database[LK14]. The same source is used to access `Amazon` user purchases, along with product tags and review meta-data, which were combined into a graph of users, bearing attributed based on the purchased product tags and review meta-data; user connections indicate purchase of a common product.

We further use publication information from the DBLP database[1] that describes a network of scientific collaborations; we recover those authors which published in any of the ICDM, SICDM or NIPS venues, tagged with their abstract and title tokens, and connected in case of a collaboration to form our `DBLP`. In our `IMDB`[2] dataset we describe the cast and crew from the filming industry connected when they contributed to the same movie. We select a list of prominent works by only focusing on the movies that have been nominated in any of 11 highly recognised festivals throughout the world (for instance the Academy Awards, Cannes, BFI, Sun-dance, Toronto, etc) and others. Features are created both by the information available for the involved individuals and by aggregating movie features (e.g,genre, year of production, region, etc).

In a smaller scale we create the `GATTWTO` dataset, based on the WTO/GATT information [Ros02]. The nodes refer to countries connected by an edge whenever a trade flow was established between them. Features are created based on country indices, as well as trade agreement memberships, resulting in an attributed multi-graph.

From the HetRec2011 workshop datasets we create the `Delicious` dataset by joining together users whenever they were linked friends in the Del.icio.us

[1] Accessed in February 2018 from `https://dblp.uni-trier.de`.
[2] Accessed in June 2018 from `https://www.imdb.com/interfaces`

Dataset	Approx. factor	Depth		
	α	$\gamma = 1/3$	$\gamma = 1/2$	$\gamma = 2/3$
Facebook	1	∞	∞	∞
Google+	0.1	∞	∞	∞
Delicious	0.3	5	5	5
Lastfm-Artists	1	∞	∞	∞
Twitter	1	∞	∞	∞
DBLP	0.3	3	3	3
IMDB	0.8	∞	6	5
GATTWTO	1	∞	∞	∞
Amazon	0.7	∞	∞	6
Lastfm-Songs	0.5	∞	∞	5

Table B.1: Dataset runtime configuration: depth and approximation factor for each used trade-off parameter γ.

social network for bookmark tracking, and assigned attributed based on the tags of their used bookmarks. From the same source we create our `Lastfm-Artists` dataset, linking users which liked the same artists, aggregating user meta-data and artist information. Finally, we access the Million Song Dataset [BEW+11] to create our `Lastfm-Songs`, which describes songs attributed with their metadata, connected between them when the Lastfm similarity score between the two exceeded the threshold of 0.3.

B.2 Kernelised Subgroup Discovery

B.2.1 Datasets

As `Chem` we refer to a dataset containing drug-like molecules as entities that are openly accessible from the *ChEMBL* [GBB+12] database, which contains several substances with their chemical and pharmaceutical properties. We derive attributes using these properties, based on the taxonomy information for the mechanism of action of each drug, the classification of the indications for which it is cleared for prescription alongside ontological information for the classification of the substance using their Experimental Factor Ontology classification [MHA+10]. The structural information that accompanies each entity is its 2D molecular structure. As a kernel we use the pre-trained similar-

ity scores from the matching entries of the PubChem infrastructure [CTP10]. This results in a dataset of 881 drugs and 882 attributes.
Further, in the `Stock` dataset we describe 2208 company stocks that are listed in the New York Stock Exchange at the time of this study. Their attribute consist of 18 traits of each company, such as the industry sector, composition of the board of directors, market information and financial indices. For each stock we collect the daily volume-weighted average prices in the year 2020. This results in a time series of 253 real values, from which we extract 1000 features using the state-of-the-art Rocket features [DPW20] for time series classification. We then train a Gaussian kernel on each feature using the method of Section 5.3.1 for a single parameter, and then train a multiple kernel of the top-performing trained kernels as per Section 5.3.2.
Finally, we use the Twitter from the SNAP [LK14] large network collection. This dataset contains ego-nets, which are a graph centred around a given user, alongside interaction between each user in these neighbours. These are, essentially, local subgraphs of the twitter interaction graph, centred around users—referred to as egos. At the same time, information of the users themselves is available in form of the recent hash-tags and followers; we use this latter information as entity attributes. For the similarity of the graphs we use the state-of-the-art Wasserstein-Weisfeiler-Lehman kernel Togninalli et al. [TGL+19], which is parametrised with a single parameter. Again, we tune this parameter using the cross validation method of Section 5.3.1.

B.2.2 Optimisation

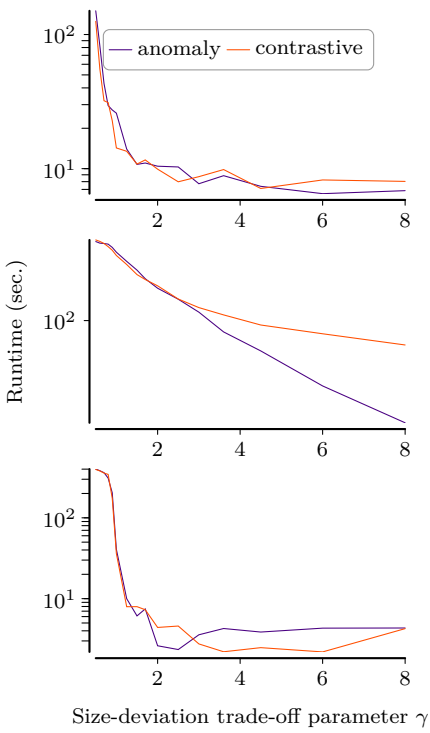
We now justify the necessity of using a branch and bound algorithm, a required part of which is the optimistic estimator we proposed in Section 5.2, and show its superiority to an equivalent problem expressed as an mixed integer quadratic program (MIQP).

The problem we seek to solve is the quantity of Eq. (5.9) over the set of describable subsets $\mathcal{L} \coloneqq \{\text{ext}(s) \mid s \subset P\}$

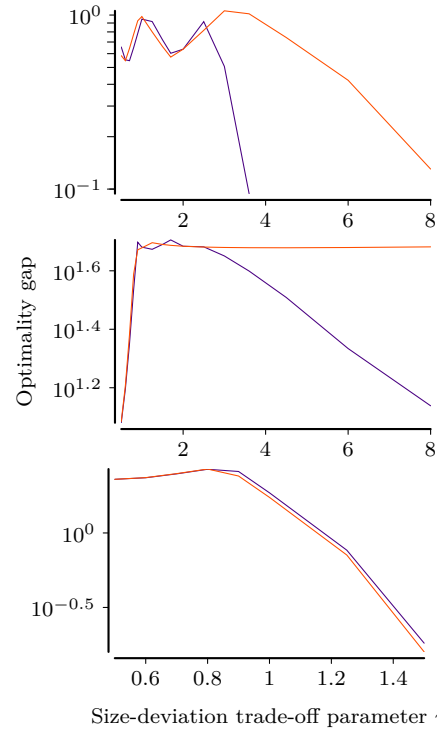
$$\max_{Q \in \mathcal{L}} a_t^{\gamma-2}(m_Q) \mathbf{z}_Q{}^\top K \mathbf{z}_Q \,, \tag{B.1}$$

with $\mathbf{z}_Q \coloneqq \mathbf{x}_Q - \frac{m_Q}{n}\mathbf{e}$, which cannot be expressed as a standard mathematical program, due to the scaling factor a_t that depends on the cardinality m_Q of Q. This dependency can only be avoided by constraining the optimisation domain over the named subsets with fixed cardinality. Then the scaling factor becomes a constant and admits a MIQP formulation.

Lemma B.1 (MIQP for Fixed Cardinality)**.** *The fixed cardinality sub-task*



(a) Runtime required for the completion of the first $d_{\max} = 3$ iterations.



(b) Optimality gap after searching all subsets $s \in \mathcal{L}$ with $|s| = 3$.

Figure B.1: Quantitative evaluation of the performance of our branch-and-bound on each of the full datasets over different values of the size-deviation trade-off parameter γ.

of our objective is equivalent to the MIQP

$$\max_{\substack{Q \in \mathcal{L} \\ |Q|=m}} \mathbf{z}_Q{}^\top K \mathbf{z}_Q \iff \tag{B.2}$$

$$
\begin{aligned}
\max_{\mathbf{z},\mathbf{x},\mathbf{p}} \quad & \mathbf{z}^\top K \mathbf{z} \\
s.t.: \quad & -x_i - \sum_{j=1}^{m} z_i(1 - v_{i,j}) \leq -1 && 1 \leq i \leq n \\
& x_i \sum_{j=1}^{m} (1 - v_{i,j}) \leq \sum_{j=1}^{m} (1 - x_i)(1 - v_{i,j}) && 1 \leq i \leq n \\
& z_i = x_i - \frac{1}{n} \sum_{j=1}^{n} x_i && 1 \leq i \leq n \\
& \sum_{i=1}^{n} x_i = m \\
& \mathbf{z} \in [-1, 1]^n, \mathbf{x} \in \{0, 1\}^n, \mathbf{p} \in \{0, 1\}^m,
\end{aligned}
\tag{B.3}
$$

for $m = |P|$ the number of predicates and $v_{i,j} \coloneqq \mathbb{1}[p_j(\epsilon_i) = \top]$ the validity of the j-th predicate on the i-th entity. The optimal subset and corresponding selector can be read from $\mathbf{x}$ and $\mathbf{p}$, which serve as characteristic vectors over the set of entities and predicates, respectively.

Proof. Let $\mathbf{x}$ and $\mathbf{p}$ be the characteristic vectors of the entities and predicates, respectively. Now any subset $Q \subset E$ and any selector $\mathcal{L} \subset P$ can be described as points in $\{0, 1\}^n \times \{0, 1\}^m$.

We first show that the first two inequalities constrain the domain of the integer lattice exactly to the points corresponding to elements of $\mathcal{L}$. The first inequality ensures that if all selected predicates are validated by an entity, it must be selected. Indeed, when all selected predicates are valid for ϵ_i, the sum in the left hand side of the equality becomes 0, so the constraint is only valid when $\mathbf{x}_Q = 1$.
The second inequality enforces that if any predicate that is invalid for ϵ_i is selected, then x_i must be 0. These two conditions describe exactly the set $\mathcal{L}$.

The third inequality ensures that the common vector $\mathbf{z}$ is equal to the corresponding $\mathbf{z}_Q$, and serves as a helper variable to formulate the objective, and the last one enforces the cardinality constraint.

The optimality of the entity subset that is induced by the optimal $\mathbf{x}$ stems from the fact that they maximise the same quantity, which is equal to that of the original problem Eq. (B.2), and the correctness of the induced selector from the correctness of the constraints. □

We can now solve this problem by solving each sub-task independently. This, however, not only requires the solution of $O(n)$ MIQP problems, but also each such problem is not solved efficiently by standard solvers, due to the specific form of the constraints.

Using our Branch-and-Bound Algorithm

We therefore use the iterative deepening depth-first search we presented in Section 2.4.2, equipped with the bound of Section 5.2.3.

To show the superiority of our method we compare it against the approach utilising a standard solver. Since the standard solver takes an excessive amount of resources, we compare the two methods on a sample from the `Stock` dataset consisting of only 100 entities. Then we invoke the GUROBI optimiser [Gur21] for each of the 100 instances of the MIQP of Lemma B.1, and show the results in Fig. B.2. It becomes clear that our algorithm outperforms a highly optimised standard solver for the proposed problem, at least in the regime of few predicates, which we deem to be the most meaningful configuration for useful descriptions.

Additionally, and to attain a more realistic evaluation, we measure the running time required for our algorithm on the entire datasets and for several values of the size-deviation trade-off parameter γ (Fig. B.1a), over all possible describable subsets with $|s| = 3$ predicates. We thus demonstrate that this method is efficient, or at least remains practical.

Optimality Guarantees

When equipped with an admissible bound, such as the one we provide in Lemma 5.2, the branch-and-bound algorithm is an exhaustive search that guarantees finding the optimal solution upon termination. Furthermore, it supports both an early termination, with a guarantee on the quality of its solution. Alternatively, it also allows for the incorporation of an approximation factor α, a user specified parameter that, when set to a value $\alpha < 1$, it performs an approximate search. In this case the found solution is satisfying the relaxed guarantee that its value is no worse than α times the true optimal.

We evaluate the optimality guarantees of our algorithm by evaluating its optimality gap, defined as the absolute difference of the best solution and the tightest bound, divided by this bound. We show in Fig. B.1b the behaviour of our algorithm over different values of the γ parameter.

Note that, for the case we only seek the best description with at most $|s|$ predicates, as in these measurements, the entire feasible set is searched

Figure B.2: Running time of the fixed cardinality MIQP sub-tasks of Lemma B.1 for at most $|s| = 3$ predicates on a small subset of 100 entities from `Stock` [left]. Invocations that exceed 1 hour are terminated prematurely (marked by red dots). For comparison the running time of our branch-and-bound algorithm on the same dataset, for up to $|s| = 5$ predicates. The latter is superior by several orders of magnitude.

and therefore the actual optimality gap is 0. For the sake of this evaluation, however, we report the optimality gap in the case no such limit would have been imposed.

B.2.3 Results

We provide the full results for the three datasets in Tables B.2 to B.4. We observe that the two methods, $t = \text{ano}$ and $t = \text{con}$ largely differ in the ranking of each subgroup, given the size of it, as expected from theory. We also see patterns in the descriptions, which vary from fine to coarse grained, e.g., in the first few subgroups of Table B.2a.

The selected entities of each named subset are also shown along the first two eigenvectors of the corresponding Gramian of each dataset in Figs. B.3 to B.5, in the same order they appear in the tabular listings. We see that as the γ parameter increases, the subsets are become either larger, in the case of $t = \text{ano}$, or closer to covering half the dataset, in the case of $t = \text{con}$.

γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[0.00 - 0.60]$	[positive_allosteric_modulator] $\wedge$ [positive_modulator]	0.003	0.2280
$[0.60 - 0.70]$	[nucleic_acid] $\wedge$ [homo_sapiens] $\wedge$ [brain_disease]	0.023	0.0628
$[0.70 - 1.00]$	[nucleic_acid] $\wedge$ [cancer] $\wedge$ [nervous_system_disease]	0.026	0.0567
$[1.00 - 1.50]$	[inhibitor] $\wedge$ [nucleic_acid] $\wedge$ [neoplasm]	0.048	0.0274
$[1.50 - 1.70]$	[inhibitor] $\wedge$ [nucleic_acid]	0.053	0.0231
$[1.70 - 2.00]$	[neoplasm] $\wedge$ [nervous_system_disease]	0.264	0.0014
$[2.00 - 2.50]$	[neoplasm]	0.411	0.0005
$[2.50 - 3.00]$	[protein] $\wedge$ [homo_sapiens]	0.784	0.0001
$[3.00 - 8.00]$	[protein]	0.938	0.0000

(a) Results for $t =$ ano

γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[0.00 - 0.60]$	[positive_allosteric_modulator] $\wedge$ [positive_modulator]	0.003	0.2296
$[0.60 - 0.70]$	[nucleic_acid] $\wedge$ [homo_sapiens] $\wedge$ [brain_disease]	0.023	0.0658
$[0.70 - 1.00]$	[nucleic_acid] $\wedge$ [cancer] $\wedge$ [nervous_system_disease]	0.026	0.0597
$[1.00 - 1.25]$	[inhibitor] $\wedge$ [nucleic_acid] $\wedge$ [neoplasm]	0.048	0.0302
$[1.25 - 1.70]$	[inhibitor] $\wedge$ [nucleic_acid]	0.053	0.0257
$[1.70 - 2.50]$	[neoplasm] $\wedge$ [nervous_system_disease]	0.264	0.0026
$[2.50 - 8.00]$	[neoplasm]	0.411	0.0014

(b) Results for $t =$ con

Table B.2: Resulting subgroups over different γ parameters of Chem.

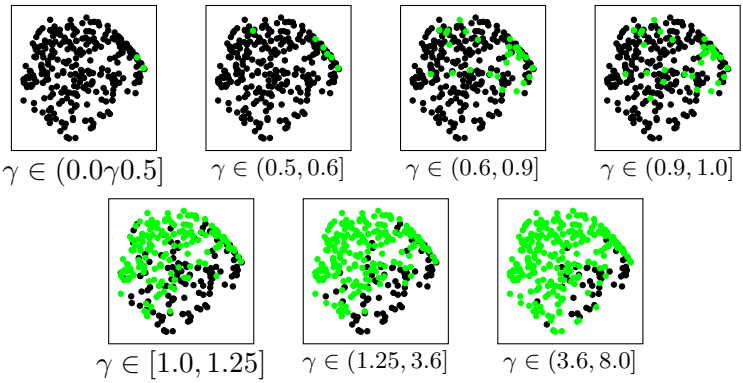
γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[0.00 - 0.50]$	$[49 \leq \texttt{price}] \wedge [\texttt{sector} = \textit{Energy}] \wedge [10 \leq \texttt{mktCap}]$	0.005	0.2767
$[0.50 - 0.60]$	$[1.9 \leq \texttt{lastDiv}] \wedge [\texttt{sector} = \textit{Energy}] \wedge [9.8 \leq \texttt{mktCap}]$	0.008	0.2315
$[0.60 - 0.90]$	$[\texttt{sector} = \textit{Energy}] \wedge [\texttt{activelyTrading}] \wedge [4.8 \leq \texttt{volAvg}]$	0.074	0.0548
$[0.90 - 1.00]$	$[\texttt{sector} = \textit{Energy}] \wedge [\texttt{activelyTrading}]$	0.082	0.0500
$[1.00 - 1.25]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}] \wedge [0.00017 \leq \texttt{lastDiv}]$	0.529	0.0064
$[1.25 - 3.60]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}]$	0.691	0.0045
$[3.60 - 8.00]$	$[10 \leq \texttt{price}]$	0.834	0.0023

(a) Results for $t = \text{ano}$

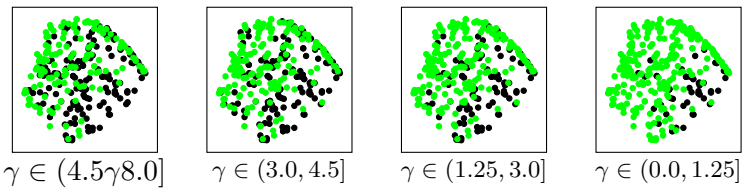
γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[4.50 - 8.00]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}] \wedge [0.00017 \leq \texttt{lastDiv}]$	0.529	0.0286
$[3.00 - 4.50]$	$[10 \leq \texttt{price}] \wedge [0.85 \leq \texttt{beta}]$	0.553	0.0298
$[1.25 - 3.00]$	$[10 \leq \texttt{price}] \wedge [0.52 \leq \texttt{beta}]$	0.691	0.0472
$[0.00 - 1.25]$	$[10 \leq \texttt{price}]$	0.834	0.0820

(b) Results for $t = \text{con}$

Table B.3: Resulting subgroups over different γ parameters of `Stock`.

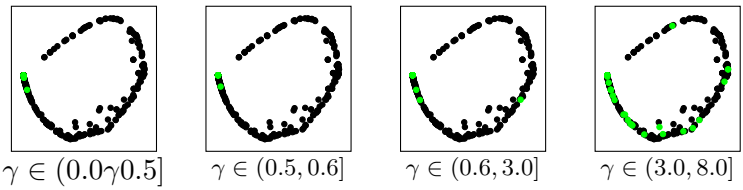


(a) Results for $t =$ ano.

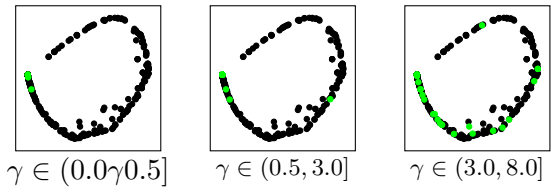


(b) Results for $t =$ con.

Figure B.3: Entities selected for each named subset found in `Stock`.

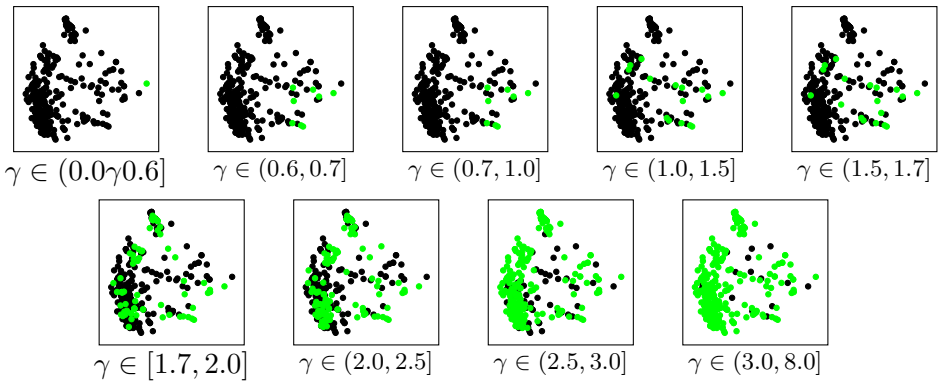


(a) Results for $t =$ ano.

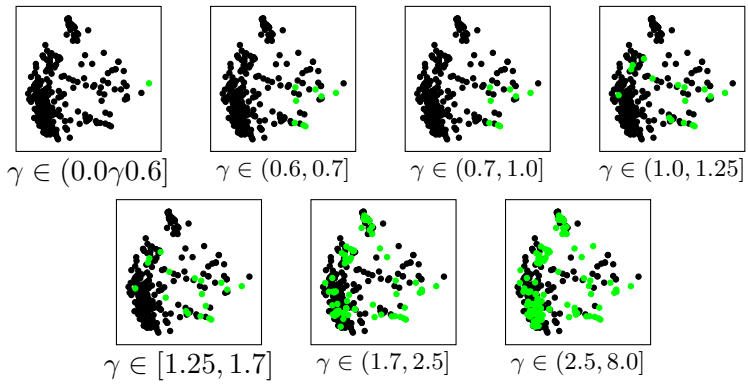


(b) Results for $t =$ con.

Figure B.4: Entities selected for each named subset found in `Twitter`.



(a) Results for $t = \text{ano}$.



(b) Results for $t = \text{con}$.

Figure B.5: Entities selected for each named subset found in `Chem`.

γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[0.00 - 0.50]$	[`@paramore:`]	0.02	0.0629
$[0.50 - 0.60]$	[`@itstayloryall`]	0.034	0.0465
$[0.60 - 3.00]$	[`@yelyahwilliams`]	0.049	0.0371
$[3.00 - 8.00]$	[`#FF`]	0.129	0.0018

(a) Results for $t = \text{ano}$

γ	**Subset Description**	$\lvert Q\rvert$	**MMD**
$[0.00 - 0.50]$	[`@paramore:`]	0.02	0.0655
$[0.50 - 3.00]$	[`@yelyahwilliams`]	0.049	0.0410
$[3.00 - 8.00]$	[`#FF`]	0.129	0.0024

(b) Results for $t = \text{con}$

Table B.4: Resulting subgroups over different γ parameters of `Twitter`.

Bibliography

[AKC+21] Quadri Adewale, Ahmed F Khan, Felix Carbonell, and Yasser Iturria-Medina. Integrated transcriptomic and neuroimaging brain model decodes biological mechanisms in aging and Alzheimer's disease. *eLife*:e62589, May 2021. DOI: 10.7554/eLife.62589 (cited on page 20).

[AV13] Abhijin Adiga and Anil Kumar S. Vullikanti. How Robust Is the Core of a Network? In *Machine Learning and Knowledge Discovery in Databases*, pages 541–556. Springer, 2013. ISBN: 978-3-642-40988-2. DOI: 10.1007/978-3-642-40988-2_35 (cited on page 136).

[ATM+12] Leman Akoglu, Hanghang Tong, Brendan Meeder, and Christos Faloutsos. PICS: Parameter-free identification of cohesive subgroups in large attributed graphs. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 439–450. SIAM, April 2012. ISBN: 978-1-61197-232-0 978-1-61197-282-5. DOI: 10.1137/1.9781611972825.38 (cited on pages 86, 96).

[AFL+11] Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sanchez, and Francisco Herrera. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 2011 (cited on page 79).

[And20] Elisha Anderson. Controversial Detroit facial recognition got him arrested for a crime he didn't commit. *Detroit Free Press*, July 2020 (cited on page 6).

[ALA+18] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*:1–78, 2018 (cited on page 8).

[ALM+16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, May 2016 (cited on page 7).

[AMP09] Andreas Argyriou, Charles A. Micchelli, and Massimiliano Pontil. When Is There a Representer Theorem? Vector Versus Matrix Regularizers. *The Journal of Machine Learning Research*:2507–2529, December 2009 (cited on page 156).

[Atz15] Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*:35–49, January 2015. DOI: 10.1002/widm.1144 (cited on pages 12, 36).

[ADM16] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*:965–984, February 2016. DOI: 10.1016/j.ins.2015.05.008 (cited on pages 14, 96, 97, 101).

[AM11] Martin Atzmueller and Folke Mitzlaff. Efficient Descriptive Community Mining. In *Twenty-Fourth International FLAIRS Conference*, March 2011 (cited on page 101).

[Bar15] Alistair Barr. Google mistakenly tags black people as 'gorillas,' showing limits of algorithms. *Wall Street Journal*, July 2015 (cited on page 6).

[BZ03] V. Batagelj and M. Zaversnik. An O(m) Algorithm for Cores Decomposition of Networks. *arXiv:cs/0310049*, October 2003 (cited on page 94).

[BCL+20] Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre, and Marc Plantevit. Identifying exceptional (dis)agreement between groups. *Data Mining and Knowledge Discovery*:394–442, March 2020. DOI: 10.1007/s10618-019-00665-9 (cited on page 11).

[BEW+11] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval*, 2011 (cited on page 180).

[Bic10] Allan Bickle. *The K-Cores of a Graph*. Western Michigan University, 2010 (cited on page 88).

[Bod21] Paula Boddington. AI and moral thinking: how can we live well with machines to enhance our moral agency? *AI and Ethics*:109–111, May 2021. DOI: 10.1007/s43681-020-00017-0 (cited on page 6).

[BGG+17] Mario Boley, Bryan R. Goldsmith, Luca M. Ghiringhelli, and Jilles Vreeken. Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery. *Data Mining and Knowledge Discovery*:1391–1418, September 2017. DOI: 10.1007/s10618-017-0520-3 (cited on pages 13, 36, 91, 92).

[BG09] Mario Boley and Henrik Grosskreutz. Non-redundant subgroup discovery using a closure system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 179–194. Springer, Springer, 2009 (cited on pages 32, 34).

[BOS+05] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*:i47–i56, June 2005. DOI: 10.1093/bioinformatics/bti1007 (cited on page 134).

[BBV21] Kailash Budhathoki, Mario Boley, and Jilles Vreeken. Discovering reliable causal rules. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 1–9. Society for Industrial and Applied Mathematics, January 2021. DOI: 10.1137/1.9781611976700.1 (cited on page 156).

[BGB+19] Antoine Buetti-Dinh, Vanni Galli, Sören Bellenberg, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V. Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, and Mark Dopson. Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports*:e00321, June 2019. DOI: 10.1016/j.btre.2019.e00321 (cited on page 6).

[CV10] Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*:277–292, September 2010. DOI: 10.1007/s10618-010-0190-x (cited on pages 64, 75).

[CŽ13] Toon Calders and Indrė Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society*, pages 43–57. Springer, 2013 (cited on page 75).

[CH08] Vira Chankong and Yacov Y Haimes. *Multiobjective Decision Making: Theory and Methodology*. Courier Dover Publications, 2008 (cited on page 59).

[Cha00] Moses Charikar. Greedy Approximation Algorithms for Finding Dense Components in a Graph. In *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 84–95. Springer-Verlag, 2000. ISBN: 978-3-540-67996-7 (cited on page 94).

[CTP10] Giovanni Cincilla, Michael Thormann, and Miquel Pons. Structuring Chemical Space: Similarity-Based Characterization of the PubChem Database. *Molecular Informatics*:37–49, January 2010. DOI: 10.1002/minf.200900015 (cited on pages 121, 181).

[Coi06] Richard Coico. Gram Staining. *Current Protocols in Microbiology*:A.3C.1–A.3C.2, 2006. DOI: 10.1002/9780471729259.mca03cs00 (cited on page 16).

[Cor21] Zoë Corbyn. Microsoft's Kate Crawford: 'AI is neither artificial nor intelligent'. *The Observer*, June 2021 (cited on page 7).

[CK21] David De Cremer and Garry Kasparov. AI Should Augment Human Intelligence, Not Replace It. *Harvard Business Review*, March 2021 (cited on page 11).

[CSE+02] Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*. MIT Press, 2002 (cited on pages 117–119).

[CRS+12] Michael J. Crowther, Richard D. Riley, Jan A. Staessen, Jiguang Wang, Francois Gueyffier, and Paul C. Lambert. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Medical Research Methodology*:34, March 2012. DOI: 10.1186/1471-2288-12-34 (cited on page 10).

[dGH07] Maria Jose del Jesus, Pedro Gonzalez, and Francisco Herrera. Multiobjective Genetic Algorithm for Extracting Subgroup Discovery Fuzzy Rules. In *2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, pages 50–57, April 2007. DOI: 10.1109/MCDM.2007.369416 (cited on page 50).

[DPW20] Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*:1454–1495, September 2020. DOI: 10.1007/s10618-020-00701-z (cited on pages 121, 181).

[DP16] Alexandra Derntl and Claudia Plant. Clustering techniques for neuroimaging applications. *WIREs Data Mining and Knowledge Discovery*:22–36, 2016. DOI: 10.1002/widm.1174 (cited on page 20).

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017 (cited on page 77).

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, second edition, 2002. DOI: 10.1017/CBO9780511755347 (cited on page 112).

[DFK16] Wouter Duivesteijn, Ad J. Feelders, and Arno Knobbe. Exceptional Model Mining: Supervised descriptive local pattern mining with complex target concepts. *Data Mining and Knowledge Discovery*:47–98, January 2016. DOI: 10.1007/s10618-015-0403-4 (cited on page 13).

[DKF+10] Wouter Duivesteijn, Arno Knobbe, Ad Feelders, and Matthijs van Leeuwen. Subgroup discovery meets Bayesian networks – An exceptional model mining approach. In *2010 IEEE International Conference on Data Mining*, pages 158–167, December 2010. DOI: 10.1109/ICDM.2010.53 (cited on page 13).

[DL19] Jannik Dunkelau and Michael Leuschel. Fairness-aware machine learning, 2019 (cited on page 64).

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012 (cited on pages 64, 75).

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*:1871–1874, June 2008 (cited on page 145).

[FFM+15] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783311 (cited on page 75).

[Fis20] Christine Fisher. Australia and the UK open joint investigation of Clearview AI, July 2020 (cited on page 6).

[FH16] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*:1–44, 2016. DOI: 10.1016/j.physrep.2016.09.002 (cited on page 96).

[GGT16] Esther Galbrun, Aristides Gionis, and Nikolaj Tatti. Top-k overlapping densest subgraphs. *Data Mining and Knowledge Discovery*, September 2016. DOI: 10.1007/s10618-016-0464-z (cited on pages 86, 88, 96, 97, 101).

[GW99] Bernhard Ganter and Rudolf Wille. Order-theoretic Foundations. In *Formal Concept Analysis: Mathematical Foundations*, pages 1–15. Springer, 1999. ISBN: 978-3-642-59830-2. DOI: 10.1007/978-3-642-59830-2_1 (cited on page 32).

[GFW03] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On Graph Kernels: Hardness Results and Efficient Alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer Berlin Heidelberg, 2003. ISBN: 978-3-540-45167-9 (cited on pages 129, 131, 133, 134, 146, 148).

[GBB+12] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*:D1100–1107, January 2012. DOI: 10.1093/nar/gkr777 (cited on pages 120, 121, 180).

[GMM+07] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*:14–es, December 2007. DOI: 10.1145/1297332.1297338 (cited on pages 64, 75).

[Gol84] A. V. Goldberg. Finding a Maximum Density Subgraph. Technical report, University of California at Berkeley, 1984 (cited on page 94).

[GBV+17] Bryan R. Goldsmith, Mario Boley, Jilles Vreeken, Matthias Scheffler, and Luca M. Ghiringhelli. Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics*:013031, January 2017. DOI: 10.1088/1367-2630/aa57c2 (cited on pages 11, 62).

[GDD06] Aurélie Goulon, Arthur Duprat, and Gérard Dreyfus. Graph Machines and Their Applications to Computer-Aided Drug Design: A New Approach to Learning from Structured Data. In *Unconventional Computation*, pages 1–19. Springer, 2006. ISBN: 978-3-540-38594-3. DOI: 10.1007/11839132_1 (cited on page 22).

[Gow71] J. C. Gower. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*:857–871, 1971. DOI: 10.2307/2528823 (cited on page 117).

[GBR+07] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007 (cited on page 112).

[Gro20] Martin Grohe. Word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data. *arXiv:2003.12590 [cs, stat]*, March 2020 (cited on page 22).

[GBK10] Henrik Grosskreutz, Mario Boley, and Maike Krause-Traudes. Subgroup discovery for election analysis: a case study in descriptive data mining. In *International Conference on Discovery Science*, pages 57–71. Springer, 2010 (cited on page 11).

[GR09] Henrik Grosskreutz and Stefan Rüping. On subgroup discovery in numerical domains. *Data Mining and Knowledge Discovery*:210–226, October 2009. DOI: 10.1007/s10618-009-0136-3 (cited on pages 13, 30, 31, 91).

[GRW08] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–456. Springer, 2008 (cited on page 43).

[GFB+10] S. Gunnemann, I. Farber, B. Boden, and T. Seidl. Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms. In *2010 IEEE International Conference on Data Mining*, pages 845–850, December 2010. DOI: 10.1109/ICDM.2010.95 (cited on page 96).

[Gur21] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2021 (cited on page 184).

[HOV+09] Sami Hanhijärvi, Markus Ojala, Niko Vuokko, Kai Puolamäki, Nikolaj Tatti, and Heikki Mannila. Tell Me Something I Don'T Know: Randomization Strategies for Iterative Data Mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–388. ACM, 2009. ISBN: 978-1-60558-495-9. DOI: 10.1145/1557019.1557065 (cited on page 75).

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised Learning. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pages 485–585. Springer, 2009. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_14 (cited on pages 6, 21).

[Hau99] David Haussler. Convolution Kernels on Discrete Structures. Technical report, Technical report, Department of Computer Science, University of California ..., 1999 (cited on page 129).

[Hel16] Sumyea Helal. Subgroup Discovery Algorithms: A Survey and Empirical Evaluation. *Journal of Computer Science and Technology*:561–576, May 2016. DOI: 10.1007/s11390-016-1647-1 (cited on pages 11, 36).

[HCG+11] Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*:495–525, December 2011. DOI: 10.1007/s10115-010-0356-2 (cited on pages 11, 13, 36).

[Ive62] Kenneth E. Iverson. *A Programming Language*. John Wiley & Sons, Inc., 1962. ISBN: 978-0-471-43014-8 (cited on page 30).

[Jan20] Darko Janjevic. Coronavirus: Gütersloh mayor Henning Schulz slams Tönnies meat producer after massive outbreak forces lockdown | DW | 24.06.2020, June 2020 (cited on page 19).

[Jay82] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*:939–952, September 1982. DOI: 10.1109/PROC.1982.12425 (cited on page 12).

[KBV17] Janis Kalofolias, Mario Boley, and Jilles Vreeken. Efficiently Discovering Locally Exceptional Yet Globally Representative Subgroups. In *2017 IEEE International Conference on Data Mining (ICDM)*, November 2017 (cited on pages 25, 26).

[KBV19] Janis Kalofolias, Mario Boley, and Jilles Vreeken. Discovering robustly connected subgraphs with simple descriptions. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1150–1155, November 2019 (cited on pages 25, 26).

[KV22] Janis Kalofolias and Jilles Vreeken. Naming the most anomalous clusters in Hilbert space for structures with attribute information. In *AAAI National Conference of the American Association for Artificial Intelligence*, 2022 (cited on pages 25, 26).

[KWV21] Janis Kalofolias, Pascal Welke, and Jilles Vreeken. SUSAN: The structural similarity random walk kernel. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 298–306. Society for Industrial and Applied Mathematics, January 2021 (cited on pages 25, 26).

[KCP10] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination Aware Decision Tree Learning. In pages 869–874. IEEE, December 2010. ISBN: 978-1-4244-9131-5. DOI: 10.1109/ICDM.2010.50 (cited on page 75).

[KAA+12] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases*:35–50, 2012 (cited on page 75).

[KTS12] U Kang, Hanghang Tong, and Jimeng Sun. Fast random walk graph kernel. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 828–838. SIAM, 2012 (cited on page 148).

[KAS+18] Navdeep Kaur, Robby Nur Aditya, Arshdeep Singh, and Tsung-Rong Kuo. Biomedical Applications for Gold Nanoclusters: Recent Developments and Future Perspectives. *Nanoscale Research Letters*:302, September 2018. DOI: 10.1186/s11671-018-2725-9 (cited on page 61).

[KKM+16] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. Benchmark data sets for graph kernels, 2016 (cited on page 142).

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016 (cited on page 75).

[Klö96] Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271, 1996 (cited on page 11).

[Klö02] Willi Klösgen. Data mining tasks and methods: Subgroup discovery: deviation analysis. In *Handbook of Data Mining and Knowledge Discovery*, pages 354–361. Oxford University Press, Inc., 2002 (cited on page 12).

[Kon80] Hiroshi Konno. Maximizing a Convex Quadratic Function Over a Hypercube. *Journal of the Operations Research Society of Japan*:171–189, 1980. DOI: 10.15807/jorsj.23.171 (cited on page 120).

[KB10] Kleanthis-Nikolaos Kontonasios and Tijl De Bie. An Information-Theoretic Approach to Finding Informative Noisy Tiles in Binary Databases. In *SDM*, pages 153–164, 2010 (cited on page 75).

[KVD11] Kleanthis-Nikolaos Kontonasios, Jilles Vreeken, and Tijl De Bie. Maximum entropy modelling for assessing results on real-valued data. In *2011 IEEE International Conference on Data Mining (ICDM)*, pages 350–359. IEEE, 2011 (cited on page 75).

[Kor85] Richard E. Korf. Depth-first Iterative-deepening: An Optimal Admissible Tree Search. *Artif. Intell.*:97–109, September 1985. DOI: 10.1016/0004-3702(85)90084-0 (cited on page 43).

[KLG+09] Petra Kralj Novak, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*:113–122, February 2009. DOI: 10.1016/j.jbi.2008.08.007 (cited on page 13).

[KJM20] Nils M. Kriege, Fredrik D. Johansson, and Christopher Morris. A survey on graph kernels:6, 2020. DOI: 10.1007/s41109-019-0195-3 (cited on pages 129, 148).

[Lai21] Nicol Turner Lee and Samantha Lai. Why New York City is cracking down on AI in hiring, December 2021 (cited on page 7).

[LCC+21] Wei Lan, Qingfeng Chen, Yi-Ping Phoebe Chen, and Wilson Wen Bin Goh. Editorial: Graph Embedding Methods for Multiple-Omics Data Analysis. *Frontiers in Genetics*:762274, 2021. DOI: 10.3389/fgene.2021.762274 (cited on page 22).

[LMA+16] Rocco Langone, Raghvendra Mall, Carlos Alzate, and Johan A. K. Suykens. Kernel Spectral Clustering and Applications. In *Unsupervised Learning Algorithms*, pages 135–161. Springer International Publishing, 2016. ISBN: 978-3-319-24211-8. DOI: 10.1007/978-3-319-24211-8_6 (cited on page 120).

[LKF+04] Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*:153–188, 2004 (cited on page 11).

[Lee20] Gavin Lee. Corona in the Slaughterhouse: The High Price of Cheap Meat. *Der Spiegel*, June 2020 (cited on page 19).

[LRJ+10] Victor E. Lee, Ning Ruan, Ruoming Jin, and Charu Aggarwal. A Survey of Algorithms for Dense Subgraph Discovery. In *Managing and Mining Graph Data*, pages 303–336. Springer US, 2010. ISBN: 978-1-4419-6045-0. DOI: 10.1007/978-1-4419-6045-0_10 (cited on page 94).

[LFK08] Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional Model Mining. In *Machine Learning and Knowledge Discovery in Databases*, pages 1–16. Springer Berlin Heidelberg, 2008. ISBN: 978-3-540-87481-2 (cited on page 13).

[LAP16] Florian Lemmerich, Martin Atzmueller, and Frank Puppe. Fast exhaustive subgroup discovery with numerical target concepts. *Data Mining and Knowledge Discovery*:711–762, 2016 (cited on page 36).

[LK14] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. June 2014 (cited on pages 121, 179, 181).

[LSL12] D. Li, X. L. Sun, and C. L. Liu. An exact solution method for unconstrained quadratic 0–1 programming: a geometric approach. *Journal of Global Optimization*:797–829, April 2012. DOI: 10.1007/s10898-011-9713-2 (cited on page 120).

[LKD+18] Jefrey Lijffijt, Bo Kang, Wouter Duivesteijn, Kai Puolamaki, Emilia Oikarinen, and Tijl De Bie. Subjectively Interesting Subgroup Discovery on Real-Valued Targets. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1352–1355, April 2018. DOI: 10.1109/ICDE.2018.00148 (cited on pages 12–14).

[Lip18] Zachary C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*:31–57, June 2018. DOI: 10.1145/3236386.3241340 (cited on page 8).

[Loa00] Charles F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*:85–100, November 2000. DOI: 10.1016/S0377-0427(00)00393-9 (cited on page 135).

[Mac21] Ryan Mac. Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men. *The New York Times*, September 2021 (cited on page 6).

[MUA+04] Pierre Mahé, Nobuhisa Ueda, Tatsuya Akutsu, Jean-Luc Perret, and Jean-Philippe Vert. Extensions of Marginalized Graph Kernels. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pages 70–. ACM, 2004. ISBN: 978-1-58113-838-2. DOI: 10.1145/1015330.1015446 (cited on page 148).

[MHA+10] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*:1112–1118, April 2010. DOI: 10.1093/bioinformatics/btq099 (cited on pages 121, 180).

[MVT12] Michael Mampaey, Jilles Vreeken, and Nikolaj Tatti. Summarizing data succinctly with the most informative itemsets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*:16, 2012 (cited on pages 64, 75).

[MK21] Marvin Meeng and Arno Knobbe. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*:158–212, January 2021. DOI: 10.1007/s10618-020-00703-x (cited on page 30).

[MS08] Kurt Mehlhorn and Peter Sanders. *Algorithms and Data Structures: The Basic Toolbox*. Springer-Verlag, 2008. ISBN: 978-3-540-77977-3 (cited on page 41).

[Mei18] Marina Meila. How to tell when a clustering is (approximately) correct using convex relaxations. *Advances in Neural Information Processing Systems*, 2018 (cited on page 120).

[Mes21] Natalia Mesa. Can the criminal justice system's artificial intelligence ever be truly fair? *Massive Science*, May 2021 (cited on page 7).

[MCC+13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013 (cited on page 22).

[MCB21] Alexandre Millot, Rémy Cazabet, and Jean-François Boulicaut. Exceptional Model Mining meets Multi-objective Optimization. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 378–386. Society for Industrial and Applied Mathematics, January 2021. DOI: 10.1137/1.9781611976700.43 (cited on page 50).

[MB14] Sandy Moens and Mario Boley. Instant exceptional model mining using weighted controlled pattern sampling. In *International Symposium on Intelligent Data Analysis*, pages 203–214. Springer, 2014 (cited on page 15).

[AH11] Awad H. Al-Mohy and Nicholas J. Higham. Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators. *SIAM Journal on Scientific Computing*:488–511, January 2011. DOI: 10.1137/100788860 (cited on page 142).

[Moo21] Mariella Moon. Facebook AI mislabels video of Black men as 'Primates' content. *Engadget*, September 2021 (cited on page 6).

[MCR+09] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. Mining Cohesive Patterns from Graphs with Feature Vectors. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 593–604. Society for Industrial and Applied Mathematics, April 2009. ISBN: 978-0-89871-682-5 978-1-61197-279-5. DOI: 10.1137/1.9781611972795.51 (cited on page 96).

[MFS+17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*:1–141, 2017. DOI: 10.1561/2200000060 (cited on page 117).

[MRS+09] Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. Subgroup Discovery for Test Selection: A Novel Approach and Its Application to Breast Cancer Diagnosis. In *Advances in Intelligent Data Analysis*

VIII, pages 119–130. Springer, 2009. ISBN: 978-3-642-03915-7. DOI: 10.1007/978-3-642-03915-7_11 (cited on page 11).

[Naj20] Alex Najibi. Racial Discrimination in Face Recognition Technology, October 2020 (cited on page 6).

[NML+18] Giannis Nikolentzos, Polykarpos Meladianos, Stratis Limnios, and Michalis Vazirgiannis. A Degeneracy Framework for Graph Similarity. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2595–2601. AAAI Press, 2018. ISBN: 978-0-9992411-2-7 (cited on pages 130, 145, 148).

[NV20] Giannis Nikolentzos and Michalis Vazirgiannis. Random Walk Graph Neural Networks. *Advances in Neural Information Processing Systems*, 2020 (cited on page 132).

[Nor20] Anna Noryskiewicz. Workers from German meat packing plant with huge coronavirus outbreak claim grim conditions. *CBS News*, June 2020 (cited on page 18).

[PMD+11] Victoria Pachón, Jacinto Mata, Juan Luis Domínguez, and Manuel J. Maña. Multi-objective Evolutionary Approach for Subgroup Discovery. In *Hybrid Artificial Intelligent Systems*, pages 271–278. Springer, 2011. ISBN: 978-3-642-21222-2. DOI: 10.1007/978-3-642-21222-2_33 (cited on page 50).

[PRT08] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568. ACM, 2008 (cited on page 75).

[PA18] Bryan Perozzi and Leman Akoglu. Discovering Communities and Anomalies in Attributed Graphs: Interactive Visual Exploration and Summarization. *ACM Trans. Knowl. Discov. Data*:24:1–24:40, January 2018. DOI: 10.1145/3139241 (cited on pages 86, 96).

[PV11] Jeff M. Phillips and Suresh Venkatasubramanian. A Gentle Introduction to the Kernel Distance. *arXiv:1103.1625 [cs]*, March 2011 (cited on page 21).

[PW20] Lucinda Platt and Ross Warwick. Are Some Ethnic Groups More Vulnerable to COVID-19 than Others? Technical report, Institute for Fiscal Studies, May 2020 (cited on page 2).

[PBV14] Simon Pool, Francesco Bonchi, and Matthijs Van Leeuwen. Description-Driven Community Detection. *ACM Transactions on Intelligent Systems and Technology*:28:1–28:28, April 2014. DOI: 10.1145/2517088 (cited on pages 86, 96).

[PGB+21] Hugo M. Proença, Peter Grünwald, Thomas Bäck, and Matthijs van Leeuwen. Discovering Outstanding Subgroup Lists for Numeric Targets Using MDL. In *Machine Learning and Knowledge Discovery in Databases*, pages 19–35. Springer International Publishing, 2021. ISBN: 978-3-030-67658-2. DOI: 10.1007/978-3-030-67658-2_2 (cited on page 13).

[Ray99] William J. Raynor. Internal disjunction. In *The International Dictionary of Artificial Intelligence*. Routledge, 1999. ISBN: 978-1-315-07410-8 (cited on page 30).

[RC19] Rebecca Reid and A. Todd Curry. The White Man Template and Academic Bias. *Inside Higher Ed*, April 2019 (cited on page 64).

[RGR17] David Reinsel, John Gantz, and John Rydning. Evolution of Data to Life-Critical. Whitepaper, Internet Data Consortium, April 2017 (cited on page 1).

[RGR18] David Reinsel, John Gantz, and John Rydning. The Digitization of the World from Edge to Core. Whitepaper, Internet Data Consortium, November 2018 (cited on page 1).

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In pages 1135–1144. ACM Press, 2016. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778 (cited on page 10).

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*:465–471, September 1978. DOI: 10.1016/0005-1098(78)90005-5 (cited on page 12).

[Ros02] Andrew K. Rose. Do We Really Know That the WTO Increases Trade? Working Paper, National Bureau of Economic Research, October 2002. DOI: 10.3386/w9273 (cited on page 179).

[Rud15] Cynthia Rudin. New models to predict recidivism could provide better way to deter repeat crime. *The Conversation*, September 2015 (cited on page 7).

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*:206–215, May 2019. DOI: 10.1038/s42256-019-0048-x (cited on page 8).

[RCC+22] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*:1–85, January 2022. DOI: 10.1214/21-SS133 (cited on page 2).

[RR19] Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, November 2019. DOI: 10.1162/99608f92.5a8a3a3d (cited on page 8).

[RWC20] Cynthia Rudin, Caroline Wang, and Beau Coker. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, March 2020. DOI: 10.1162/99608f92.6ed64b30 (cited on page 7).

[SS16] Saburou Saitoh and Yoshihiro Sawano. Fundamental Properties of RKHS. In *Theory of Reproducing Kernels and Applications*, pages 65–160. Springer, 2016. ISBN: 978-981-10-0530-5. DOI: 10.1007/978-981-10-0530-5_2 (cited on page 111).

[Sch00] Bernhard Schölkopf. The kernel trick for distances. In *Advances in Neural Information Processing Systems*. MIT Press, 2000 (cited on page 129).

[SBD+22] Rianne M. Schouten, Marcos L. P. Bueno, Wouter Duivesteijn, and Mykola Pechenizkiy. Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery*:379–413, January 2022. DOI: 10.1007/s10618-021-00808-x (cited on page 13).

[SHW+21] Till Hendrik Schulz, Tamás Horváth, Pascal Welke, and Stefan Wrobel. A Generalized Weisfeiler-Lehman Graph Kernel, January 2021 (cited on page 130).

[Sei83] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*:269–287, September 1983. DOI: 10.1016/0378-8733(83)90028-X (cited on page 136).

[SEF16] Kijung Shin, Tina Eliassi-Rad, and Christos Faloutsos. CoreScope: Graph Mining Using k-Core Analysis—Patterns, Anomalies and Algorithms. In *IEEE International Conference on Data Mining*, pages 469–478. IEEE, 2016 (cited on pages 88, 89, 94).

[SJX+17] null Shirui Pan, null Jia Wu, null Xingquan Zhu, null Guodong Long, and null Chengqi Zhang. Task Sensitive Feature Exploration and Learning for Multitask Graph Classification. *IEEE transactions on cybernetics*:744–758, March 2017. DOI: 10.1109/TCYB.2016.2526058 (cited on pages 131, 142).

[Sho87] N. Z. Shor. Class of global minimum bounds of polynomial functions. *Cybernetics*:731–734, November 1987. DOI: 10.1007/BF01070233 (cited on page 120).

[SMZ12] Arlei Silva, Wagner Meira Jr., and Mohammed J. Zaki. Mining Attribute-structure Correlated Patterns in Large Attributed Graphs. *Proc. VLDB Endow.*:466–477, January 2012. DOI: 10.14778/2140436.2140443 (cited on pages 96, 97, 99, 101).

[Sim18] Tom Simonite. When It comes to Gorillas, Google Photos remains blind. *Wired*, January 2018 (cited on page 6).

[SCK90] R. K. Sizemore, J. J. Caldwell, and A. S. Kendrick. Alternate gram staining technique using a fluorescent lectin. *Applied and Environmental Microbiology*:2245–2247, July 1990. DOI: 10.1128/aem.56.7.2245-2247.1990 (cited on pages 15, 16).

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems 25*:2951–2959, 2012 (cited on page 121).

[Sof16] Kim Soffen. How racial gerrymandering deprives black people of political power. *Washington Post*, June 2016 (cited on page 64).

[SFK15] Hao Song, Peter Flach, and Georgios Kalogridis. Dataset Shift Detection with Model-Based Subgroup Discovery. In *2nd International Workshop on Learning over Multiple Contexts (LMCE 2015)*, page 10, September 2015 (cited on page 13).

[SKF+16] Hao Song, Meelis Kull, Peter Flach, and Georgios Kalogridis. Subgroup Discovery with Proper Scoring Rules. In *Machine Learning and Knowledge Discovery in Databases*, pages 492–510. Springer, Cham, September 2016. ISBN: 978-3-319-46226-4. DOI: 10.1007/978-3-319-46227-1_31 (cited on pages 13, 36, 37).

[SCR04] Arnaud Soulet, Bruno Crémilleux, and François Rioult. Condensed Representation of Emerging Patterns. In *Advances in Knowledge Discovery and Data Mining*, pages 127–132. Springer, 2004. ISBN: 978-3-540-24775-3. DOI: 10.1007/978-3-540-24775-3_16 (cited on page 50).

[SR11] Sujatha Srinivasan and Sivakumar Ramakrishnan. Evolutionary multi objective optimization for rule mining: a review. *Artificial Intelligence Review*:205, March 2011. DOI: 10.1007/s10462-011-9212-3 (cited on page 50).

[Ste17a] Roman Steinberg. 6 areas where artificial neural networks outperform humans, December 2017 (cited on page 6).

[Ste17b] Nicholas O. Stephanopoulos. The Causes and Consequences of Gerrymandering. *William & Mary Law Review*:2115, 2017 (cited on page 64).

[SBG+20] Christopher Sutton, Mario Boley, Luca M. Ghiringhelli, Matthias Rupp, Jilles Vreeken, and Matthias Scheffler. Identifying domains of applicability of machine learning models for materials science. *Nature Communications*:4428, September 2020. DOI: 10.1038/s41467-020-17112-9 (cited on page 11).

[AHM+21] Zainab Al-Taie, Mark Hannink, Jonathan Mitchem, Christos Papageorgiou, and Chi-Ren Shyu. Drug Repositioning and Subgroup Discovery for Precision Medicine Implementation in Triple Negative Breast Cancer. *Cancers*:6278, December 2021. DOI: 10.3390/cancers13246278 (cited on page 11).

[Tan58] T. T. Tanimoto. *An Elementary Mathematical Theory of Classification and Prediction by T.T. Tanimoto*. International Business Machines Corporation New York, 1958 (cited on pages 116, 117).

[Tat08] Nikolaj Tatti. Maximum entropy based significance of itemsets. *Knowledge and Information Systems*:57–77, October 2008. DOI: 10.1007/s10115-008-0128-4 (cited on page 75).

[19] The System Is Rigged: Student Debt and the Racial Wealth Gap. Press Release, Roosevelt Institute, September 2019 (cited on page 64).

[TFL00] Ljupčo Todorovski, Peter Flach, and Nada Lavrač. Predictive Performance of Weighted Relative Accuracy. In *Principles of Data Mining and Knowledge Discovery*, pages 255–264. Springer, 2000. ISBN: 978-3-540-45372-7. DOI: 10.1007/3-540-45372-5_25 (cited on page 35).

[TGL+19] Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein Weisfeiler-Lehman Graph Kernels. *Advances in Neural Information Processing Systems*, 2019 (cited on pages 23, 121, 130, 145, 181).

[TLT08] Igor Trajkovski, Nada Lavrač, and Jakub Tolar. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*:588–601, August 2008. DOI: 10.1016/j.jbi.2007.12.001 (cited on page 36).

[Tso15] Charalampos Tsourakakis. The K-clique Densest Subgraph Problem. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1122–1132. International World Wide Web Conferences Steering Committee, 2015. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741098 (cited on page 94).

[TBG+13] Charalampos E Tsourakakis, Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Maria A Tsiarli. Denser than the Densest Subgraph: Extracting Optimal Quasi-Cliques with Quality Guarantees. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 104–112. ACM, 2013. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487645 (cited on page 94).

[Tso14] Charalampos E. Tsourakakis. A Novel Approach to Finding Near-Cliques: The Triangle-Densest Subgraph Problem. *arXiv:1405.147[cs]*, 2014 (cited on page 94).

[UBC+17] Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems. *Artificial Intelligence*:48–69, March 2017. DOI: 10.1016/j.artint.2015.04.003 (cited on page 50).

[UAU+03] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets. In *Workshop on Frequent Itemset Mining Implementations*, 2003 (cited on page 42).

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*:2579–2605, 2008 (cited on page 22).

[vLee10] Matthijs van Leeuwen. Maximal exceptions with minimal descriptions. *Data Mining and Knowledge Discovery*:259–276, September 2010. DOI: 10.1007/s10618-010-0187-5 (cited on page 13).

[vLK11] Matthijs van Leeuwen and Arno Knobbe. Non-redundant subgroup discovery in large and complex data. In *Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer, 2011. ISBN: 978-3-642-23808-6. DOI: 10.1007/978-3-642-23808-6_30 (cited on page 49).

[vLK12] Matthijs van Leeuwen and Arno Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*:208–242, September 2012. DOI: 10.1007/s10618-012-0273-y (cited on pages 12, 13).

[vLU13] Matthijs van Leeuwen and Antti Ukkonen. Discovering Skylines of Subgroup Sets. In *Machine Learning and Knowledge Discovery in Databases*, pages 272–287. Springer, 2013. ISBN: 978-3-642-40994-3. DOI: 10.1007/978-3-642-40994-3_18 (cited on page 50).

[VBS06] S. V. N. Vishwanathan, Karsten M. Borgwardt, and Nicol N. Schraudolph. Fast computation of graph kernels. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 1449–1456. MIT Press, 2006 (cited on pages 134, 148).

[VSK+10] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph Kernels. *Journal of Machine Learning Research*:1201–1242, 2010 (cited on pages 22, 129–131, 134, 135, 142, 143, 148).

[VLS11] Jilles Vreeken, Matthijs Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*:169–214, July 2011. DOI: 10.1007/s10618-010-0202-x (cited on page 12).

[Wae22] Rosalie A. Waelen. The struggle for recognition in the age of facial recognition technology. *AI and Ethics*, March 2022. DOI: 10.1007/s43681-022-00146-8 (cited on page 6).

[WZD+18] Shui-Hua Wang, Yu-Dong Zhang, Zhengchao Dong, and Preetha Phillips. Neuroimaging modalities. In *Pathological Brain Detection*, pages 13–28. Springer Singapore, 2018. ISBN: 978-981-10-4026-9. DOI: 10.1007/978-981-10-4026-9âĆĆ (cited on page 20).

[Web01] Geoffrey I. Webb. Discovering Associations with Numeric Variables. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 383–388. ACM, 2001. ISBN: 1-58113-391-X. DOI: 10.1145/502512.502569 (cited on pages 11, 35).

[Wei62] Paul M Weichsel. The Kronecker product of graphs. *Proceedings of the American mathematical society*:47–52, 1962 (cited on page 133).

[Wet21] Nicole Wetsman. WHO outlines principles for ethics in health AI. *The Verge*, June 2021 (cited on page 7).

[Wro97] Stefan Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer-Verlag, 1997. ISBN: 978-3-540-63223-8 (cited on page 11).

[WKR+08] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*:1–37, January 2008. DOI: 10.1007/s10115-007-0114-2 (cited on page 30).

[YFS+20] Fang Yang, Kunjie Fan, Dandan Song, and Huakang Lin. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics*:323, July 2020. DOI: 10.1186/s12859-020-03646-8 (cited on page 22).

[Yüc02] Ümit Yüceer. Discrete convexity: convexity for functions defined on discrete spaces. *Discrete Applied Mathematics*:297–304, July 2002. DOI: 10.1016/S0166-218X(01)00191-3 (cited on pages 71, 166).

[ZP20] Hugo Zeberg and Svante Pääbo. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*:610–612, November 2020. DOI: 10.1038/s41586-020-2818-3 (cited on page 2).

[ZWS+13] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning Fair Representations. *ICML (3)*:325–333, 2013 (cited on pages 64, 75).

[ZSY+21] Bingxin Zhao, Yue Shan, Yue Yang, Zhaolong Yu, Tengfei Li, Xifeng Wang, Tianyou Luo, Ziliang Zhu, Patrick Sullivan, Hongyu Zhao, Yun Li, and Hongtu Zhu. Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits. *Nature Communications*:2878, May 2021. DOI: 10.1038/s41467-021-23130-y (cited on page 20).

Alphabetical Index