2016•2017
FACULTEIT WETENSCHAPPEN
master in de informatica

Masterproef

Parallel-correctness and transferability for conjunctive queries under bag semantics

Promotor :
Prof. dr. Frank NEVEN

Brecht Vandevoort

Scriptie ingediend tot het behalen van de graad van master in de informatica

2016•2017
FACULTEIT WETENSCHAPPEN
master in de informatica

Masterproef

Parallel-correctness and transferability for conjunctive queries under bag semantics

Promotor :
Prof. dr. Frank NEVEN

Brecht Vandevoort
Scriptie ingediend tot het behalen van de graad van master in de informatica

universiteit hasselt | Maastricht University

Abstract

Due to the increasing popularity of cloud computing and big data, there is a growing need for data processing in distributed and parallel settings. Of particular interest are the evaluations of queries in a single-round of communication where data is distributed over different servers according to some distribution policy, after which each server evaluates the query over the locally available data. Based on this setting, Ameloot et al. [4] introduced a correctness condition, called parallel-correctness, and studied this condition as well as transferability of parallel-correctness while considering unions of conjunctive queries under set semantics. In this thesis, we extend this study toward bag semantics, as bag semantics are often used in practice and their usage is inevitable for certain aggregation functions. We provide characterizations for both parallel-correctness and transferability for conjunctive queries with inequalities under bag semantics and use these characterizations to study the complexity of these problems. The existing distributed evaluation model is however quite restrictive on possible distribution policies for certain conjunctive queries under bag semantics. We therefore propose a slightly modified model based on ordered networks.

Acknowledgements

I would like to express my special thanks of gratitude to my promotor, Prof. dr. Frank Neven, and my supervisor, Bas Ketsman, for introducing me to this topic and for their helpful thoughts and remarks on my work, allowing me to further refine this thesis.

Secondly, I would like to thank my girlfriend, my parents, my family and friends for their support and their encouragements along the way.

Dutch summary
Nederlandse samenvatting

Inleiding

Vanwege de groeiende populariteit van cloud computing en big data is er steeds meer nood aan het verwerken van zeer grote hoeveelheden data. Voor deze toepassingen wordt typisch gebruik gemaakt van gedistribueerde systemen bestaande uit honderden tot zelfs duizenden servers waarbij elke server een deel van de data toegewezen krijgt. Het MapReduce model [10], ontwikkeld door Google, vereenvoudigt het efficiënt verwerken van deze data over een netwerk van servers door middel van abstractie. Een gebruiker van het MapReduce model moet bijvoorbeeld niet zelf communicatie tussen de verschillende servers implementeren of hardware storingen afhandelen, aangezien het MapReduce model dit onderliggend reeds voorziet. Recente technologieën zoals Spark [5] combineren dit MapReduce model met in-memory systemen. Het voordeel van deze in-memory systemen is dat data niet langer geladen moet worden vanuit extern geheugen zoals bijvoorbeeld een harde schijf, waardoor de verwerking ervan sneller is. In tegenstelling tot traditionele databases wordt de uitvoeringstijd nu niet langer bepaald door het aantal IO-operaties, maar door de hoeveelheid netwerkcommunicatie tussen de verschillende servers.

Gebaseerd op deze gedistribueerde systemen introduceerden Koutris en Suciu [14] een massively parallel communication (MPC) model bestaande uit een netwerk van servers, nodes genaamd, gebruik makend van een shared-nothing architectuur. In dit MPC model worden berekeningen uitgevoerd door afwisselende fases van enerzijds globale synchronisatie en communicatie en anderzijds berekeningen per node. Tijdens de communicatie fase kunnen de verschillende nodes data uitwisselen, waarna elke node in de daaropvolgende fase berekeningen doet op de lokaal beschikbare data. Dit MPC model maakt het mogelijk om de complexiteit van algoritmes in gedistribueerde omgevingen te bestuderen en vergelijken. Zo onderzochten Beame, Koutris en Suciu [6, 14] de complexiteit van het evalueren van conjunctieve queries in dit MPC model.

Een bijzonder geval van dit MPC model zijn single-round evaluaties van queries. In dit geval wordt de data eerst gedistribueerd aan de hand van een distribution policy, waarna elke node de query uitvoert op de lokale data. Het uiteindelijke resultaat wordt bekomen door de unie te nemen over al deze lokale resultaten. Een specifieke familie van distribution policies die bruikbaar zijn bij de evaluatie van conjunctieve queries in het single-round MPC model zijn Hypercube distributions [7]. Een Hypercube distribution verdeelt de data over de nodes op basis van de structuur van de conjunctieve query. Deze techniek is reeds terug te vinden in het werk van Ganguly, Silberschatz en Tsur [11], en is door Afrati en Ullman [3] bestudeerd in de context van MapReduce.

Ameloot et al. [4] beschreven een algemeen framework dat toelaat single-round evaluaties van conjunctieve queries onder willekeurige distribution policies te bestuderen. Meer bepaald introduceerden ze de volgende twee eigenschappen voor queries en distribution policies:

- *Parallel-correctness:* Gegeven een query $\mathcal{Q}$ en een distribution policy $\boldsymbol{P}$, produceert de single-round gedistribueerde evaluatie van $\mathcal{Q}$ volgens $\boldsymbol{P}$ steeds het correcte resultaat, onafhankelijk van de data waarover $\mathcal{Q}$ geëvalueerd wordt?

- *Parallel-correctness transfer:* Gegeven twee queries $\mathcal{Q}$ en $\mathcal{Q}'$, is $\mathcal{Q}'$ parallel-correct onder elke distribution policy $\boldsymbol{P}$ waaronder $\mathcal{Q}$ parallel-correct is?

Deze tweede eigenschap laat toe om meerdere queries achter elkaar uit te voeren zonder de data opnieuw te distribueren. Deze eigenschap is daarom voornamelijk nuttig in een setting met geautomatiseerde datadistributie waarbij het doel is om de data optimaal te verdelen wanneer meerdere queries geëvalueerd moeten worden. Ameloot et al. [4] onderzochten vervolgens deze eigenschappen in het kader van unies van conjunctieve queries met ongelijkheden. Voor beide eigenschappen beschreven ze een karakterisering die hen toeliet om de complexiteit van deze eigenschappen te bepalen. Meer specifiek toonden ze aan dat parallel-correctness beslissen Π_2^p-compleet is, zelfs voor conjunctieve queries zonder unies en ongelijkheden. Parallel-correctness transfer beslissen is daarentegen Π_3^p-compleet.

De resultaten van Ameloot et al. [4] zijn gebaseerd op conjunctieve queries onder set semantiek, wat inhoudt dat mogelijke duplicaten in het resultaat genegeerd worden. In de praktijk worden queries echter vaak onder bag semantiek geëvalueerd, wat inhoudt dat mogelijke duplicaten niet verwijderd worden uit het resultaat, tenzij expliciet aangegeven. De reden hiervoor is dat het verwijderen van duplicaten mogelijk veel tijd inneemt bij grotere datasets. Bovendien is het behoud van deze duplicaten essentieel voor aggregatie functies waarbij bijvoorbeeld het aantal voorkomens van elk resultaat geteld wordt. Chaudhuri en Vardi [8] definieerden reeds

de evaluatie van conjunctieve queries onder bag semantiek en onderzochten optimalisatie en containment in deze setting.

Vanwege het praktisch nut van evaluaties van queries onder bag semantiek breiden we in deze thesis het werk van Ameloot et al. [4] uit naar bag semantiek. We voorzien eerst alternatieve karakteriseringen voor parallel-correctness en transferability onder bag semantiek, waarna we deze karakteriseringen gebruiken om de complexiteit van beide eigenschappen te bestuderen. Vervolgens wordt voor beide eigenschappen het verband tussen set en bag semantiek bestudeerd. Aangezien parallel-correctness onder bag semantiek zeer restrictief blijkt te zijn voor toegelaten distribution policies, stellen we een alternatief gedistribueerd model voor gebaseerd op geordende netwerken.

Terminologie

Deze sectie geeft een informele samenvatting van de gebruikte terminologie. Voor de formele definities verwijzen we naar Chapter 2.

Een *instance* onder set semantiek is een verzameling van facts. Onder bag semantiek is een instance een verzameling van annotated facts. Deze annotated facts zijn combinaties van een fact en een multipliciteit, waarbij deze multipliciteit intuïtief aangeeft hoeveel keer een bepaalde fact voorkomt in een instance.

Een *conjunctieve query met ongelijkheden* $\mathcal{Q}$ is van de vorm

$$T(\mathbf{x}) \leftarrow R_1(\mathbf{y_1}), \ldots, R_m(\mathbf{y_m}), \beta_1, \ldots, \beta_p$$

waarbij elke R_i de naam van een relatie is en elke $\mathbf{y_i}$ een verzameling van variabelen. We verwijzen naar de atomen $R_i(\mathbf{y_i})$ in $\mathcal{Q}$ als $body_\mathcal{Q}$ en naar $T(\mathbf{x})$ als $head_\mathcal{Q}$. Bij deze laatste stelt $\mathbf{x}$ een verzameling van variabelen voor die elk in een $\mathbf{y_i}$ voorkomen. Elke β_i stelt een ongelijkheid $z \neq z'$ voor waarbij z en z' variabelen zijn die ook in een $\mathbf{y_i}$ voorkomen. We verwijzen naar de verzameling van alle conjunctieve queries met ongelijkheden als $\mathbf{CQ}^{\neq}$ en naar de verzameling van alle conjunctieve queries zonder ongelijkheden als $\mathbf{CQ}$. Merk op dat per definitie $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$.

Een *valuatie* V voor een conjunctieve query met ongelijkheden $\mathcal{Q}$ is een totale functie die variabelen in $\mathcal{Q}$ afbeeldt op waarden uit een universum U en die bovendien de ongelijkheden in $\mathcal{Q}$ respecteert. De facts $V(body_\mathcal{Q})$ zijn de benodigde facts voor een valuatie V. Indien een instance I al deze facts bevat, leidt V het fact $V(head_\mathcal{Q})$ af. Onder bag semantiek wordt dit afgeleide fact bovendien gecombineerd met een multipliciteit die bepaald wordt door de multipliciteiten van de benodigde facts voor V in I. Het uiteindelijke resultaat $\mathcal{Q}(I)$ van een query $\mathcal{Q}$ op een instance I is de unie van alle afgeleide facts door valuaties voor $\mathcal{Q}$ waarvoor de benodigde facts in I aanwezig zijn.

Een netwerk bestaat uit een verzameling van nodes. Facts worden over deze nodes gedistribueerd aan de hand van een *distribution policy* $\boldsymbol{P}$. We gebruiken de notatie $\mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ om de verzameling van facts te beschrijven die volgens $\boldsymbol{P}$ op een node κ gemapt worden. Merk in het bijzonder op dat facts naar een willekeurig aantal nodes gemapt kunnen worden en dat de voorafgaande verdeling van facts over verschillende nodes geen rol speelt in de distributie van deze facts. De single-round evaluatie van een conjunctieve query $\mathcal{Q}$ over een netwerk bestaat uit twee stappen. Eerst wordt $\mathcal{Q}$ op elke node apart geëvalueerd, waarna het uiteindelijke resultaat bepaald wordt door de unie van al deze lokale resultaten te nemen. We noteren de gedistribueerde evaluatie van een query $\mathcal{Q}$ over een instance I aan de hand van een distribution policy $\boldsymbol{P}$ als $[\mathcal{Q}, \boldsymbol{P}](I)$.

Parallel-correctness en transferability

Een query $\mathcal{Q}$ is *parallel-correct* onder een distribution policy $\boldsymbol{P}$ indien voor alle instances I geldt dat $\mathcal{Q}(I) = [\mathcal{Q}, \boldsymbol{P}](I)$. De gedistribueerde evaluatie van $\mathcal{Q}$ volgens $\boldsymbol{P}$ moet met andere woorden steeds het correcte resultaat opleveren. In deze thesis onderzoeken we deze eigenschap voor conjunctieve queries met ongelijkheden onder bag semantiek. We tonen aan dat een query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ parallel-correct is onder een distribution policy $\boldsymbol{P}$ over een netwerk $\mathcal{N}$ onder bag semantiek als en slechts als voor elke valuatie V voor $\mathcal{Q}$ er exact één node κ in $\mathcal{N}$ bestaat zodat $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. Met andere woorden moet elke valuatie V die gebruikt wordt in de evaluatie van $\mathcal{Q}(I)$ over een instance I ook in de gedistribueerde omgeving exact één keer toegepast worden. Dit kan intuïtief ingezien worden door het feit dat het ontbreken van valuaties leidt tot ontbrekende resultaten, terwijl het dubbel gebruiken van valuaties leidt tot dubbele resultaten.

Aan de hand van deze karakterisering tonen we vervolgens aan dat het bepalen of een query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ parallel-correct is onder een distribution policy $\boldsymbol{P}$ onder bag semantiek in Π_2^p zit. Deze bonvengrens kan echter verbeterd worden indien we ons beperken tot distribution policies waarbij de toekenning van facts aan nodes beschreven kan worden aan de hand van een deterministisch polynomiaal algoritme. In dat geval zit het beslissen van parallel-correctness meer bepaald in coNP. We tonen bovendien aan dat dit probleem coNP-compleet is. Deze bovengrens is met andere woorden ook onmiddellijk een ondergrens voor parallel-correctness van conjunctieve queries onder bag semantiek. Het bewijs van deze ondergrens wordt geleverd via een reductie van 3-SAT, een welbekend NP-compleet probleem [9, 13], naar het complement van het parallel-correctness probleem.

Beschouw twee queries $\mathcal{Q}$ en $\mathcal{Q}'$. We zeggen dat er *parallel-correctness transfer* is van $\mathcal{Q}$ naar $\mathcal{Q}'$ indien $\mathcal{Q}'$ parallel-correct is onder elke distribution

policy waaronder $\mathcal{Q}$ parallel-correct is. Zoals reeds aangehaald in de inleiding is deze eigenschap zeer nuttig indien meerdere queries geëvalueerd moeten worden, aangezien het toelaat meerdere queries uit te voeren zonder de data opnieuw te moeten distribueren na elke query.

De bovenstaande karakterisering van parallel-correctness onder bag semantiek impliceert onrechtstreeks dat, afhankelijk van de query $\mathcal{Q}$ in kwestie, de benodigde facts van bepaalde valuaties voor $\mathcal{Q}$ steeds samen gegroepeerd moeten worden op een willekeurige node in het netwerk. Dit is een gevolg van het feit dat de benodigde facts van elke valuatie slechts op één node mogen samenkomen, gecombineerd met de observatie dat de benodigde facts van bepaalde valuaties strikte subsets zijn van de benodigde facts van andere valuaties. We gebruiken $\mathit{impFacts}(V, \mathcal{Q})$ om deze verzameling van alle facts te beschrijven die steeds voorkomen op een node waarop de benodigde facts van een valuatie V voor $\mathcal{Q}$ voorkomen onder distribution policies waaronder $\mathcal{Q}$ parallel-correct is. In deze thesis wordt een verzameling van afleidingsregels voor $\mathit{impFacts}(V, \mathcal{Q})$ beschreven die bovendien compleet zijn indien we werken onder een eindig domein van mogelijke waarden.

Onder bag semantiek is er parallel-correctness transfer van een query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ naar een query $\mathcal{Q}' \in \mathbf{CQ}^{\neq}$ als en slechts als voor elke valuatie V' voor $\mathcal{Q}'$ er een valuatie V voor $\mathcal{Q}$ is zodat $V(\mathit{body}_{\mathcal{Q}}) \subseteq V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V, \mathcal{Q})$. Deze karakterisering leidt tot de observatie dat het beslissen van transferability voor queries in $\mathbf{CQ}^{\neq}$ onder bag semantiek in EXPTIME zit. Deze bovengrens wordt bepaald door de berekening van $\mathit{impFacts}(V, \mathcal{Q})$ voor elke valuatie V voor $\mathcal{Q}$ over een eindig domein. Daarom kan deze bovengrens verlaagd worden tot Π_2^p indien we ons beperken tot conjunctieve queries met ongelijkheden zonder self-joins. Voor deze conjunctieve queries zonder self-joins geldt immers steeds dat $\mathit{impFacts}(V, \mathcal{Q}) = V(\mathit{body}_{\mathcal{Q}})$.

Vergelijking tussen set en bag semantiek

Ameloot et al. [4] toonden reeds aan dat de karakteriseringen van parallel-correctness en transferability voor conjunctieve queries onder set semantiek gerelateerd zijn aan de notie van minimale valuaties. Deze minimale valuaties zijn valuaties waarvoor er geen andere valuaties bestaan die hetzelfde fact afleiden maar strikt minder facts vereisen. Onder bag semantiek is deze notie niet langer van belang, aangezien elke valuatie van belang is om een correcte multipliciteit te bepalen. Ondanks dit verschil blijken er toch nog bepaalde relaties tussen set en bag semantiek te zijn.

Voor parallel-correctness blijkt meer bepaald dat een conjunctieve query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ steeds parallel-correct is onder een distribution policy $\boldsymbol{P}$ onder set semantiek indien deze query $\mathcal{Q}$ parallel-correct is onder $\boldsymbol{P}$ onder bag semantiek. In de omgekeerde richting is dit helaas niet altijd het geval. Wanneer we ons echter beperken tot strongly minimal conjunctieve queries en

nonreplicating distribution policies vallen parallel-correctness onder set en bag semantiek samen. Deze strongly minimal queries zijn kort samengevat queries waarbij elke valuatie minimaal is. Nonreplicating distribution policies zijn anderzijds distribution policies waarbij elk fact op hoogstens één node gemapt wordt.

Voor transferability blijkt er helaas geen onmiddellijk verband te zijn in het algemeen. In het bijzonder impliceert transferability onder bag semantiek niet noodzakelijk transferability onder set semantiek, ondanks het feit dat parallel-correctness onder bag semantiek wel steeds parallel-correctness onder set semantiek impliceert. Indien we ons opnieuw beperken tot strongly minimal conjunctieve queries en nonreplicating distribution policies vallen parallel-correctness transfer voor set en bag semantiek wel opnieuw samen.

Geordende netwerken

Hierboven werd reeds kort aangehaald dat de karakterisering van parallel-correctness onder bag semantiek onrechtstreeks een bepaalde groepering van facts op de verschillende nodes impliceert. Afhankelijk van de query $\mathcal{Q}$ kan dit zelfs betekenen dat er geen distribution policy $\boldsymbol{P}$ bestaat zodat $\mathcal{Q}$ parallel-correct is onder $\boldsymbol{P}$ en $\boldsymbol{P}$ de facts zodanig verspreid dat niet alle valuaties op dezelfde node worden afgeleid. Voor dergelijke queries is de beschreven gedistribueerde omgeving volgens het MPC model compleet nutteloos, aangezien bij deze queries niet efficiënt gebruik gemaakt kan worden van de verschillende nodes om de data te verdelen.

Omwille van deze beperkingen stellen we een aanpassing aan het gedistribueerd model voor die de karakterisering van parallel-correctness onder bag semantiek minder strikt maakt. Deze aanpassing voegt meer bepaald een totale orde toe over de nodes in het netwerk. Een node κ in het geordend netwerk leidt enkel een fact af op basis van een bepaalde valuatie V indien deze node alle benodigde facts bevat en er geen andere node κ' in het netwerk bestaat die ook al de benodigde facts voor V bezit en volgens de orde vóór κ komt.

Merk op dat deze aanpassing het resultaat onder set semantiek niet aanpast, waardoor de karakteriseringen voor parallel-correctness en transferability onder set semantiek behouden blijven. Onder bag semantiek leiden deze aanpassingen echter wel tot een aangepast resultaat. Geordende netwerken laten in het bijzonder toe dat onder bag semantiek meerdere nodes de benodigde facts voor een bepaalde valuatie mogen bevatten, aangezien van deze nodes toch enkel de eerste deze valuatie effectief zal gebruiken om een fact af te leiden. Met andere woorden wordt de karakterisering van parallel-correctness onder bag semantiek onder geordende netwerken als volgt vereenvoudigd: een query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is parallel-correct onder een distribution policy $\boldsymbol{P}$ over een geordend netwerk $\mathcal{N}$ als en slechts als voor elke valu-

atie V voor $\mathcal{Q}$ er een node $\kappa \in \mathcal{N}$ bestaat zodat $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. De bepaalde bovengrenzen op de complexiteit van parallel-correctness onder bag semantiek blijven ongewijzigd onder deze aanpassing. De karakterisering voor transferability vereenvoudigt aanzienlijk, waardoor de bovengrens op de complexiteit verbeterd kan worden tot Π_2^p.

Hypercube distributions

Hypercube distributions voor een conjunctieve query $\mathcal{Q}$ zijn distribution policies die de data verdelen over de verschillende nodes op basis van de structuur van $\mathcal{Q}$. Meer bepaald organiseert een Hypercube distribution de nodes in het netwerk volgens een hypercube met een dimensie per variabele in $\mathcal{Q}$.

Ameloot et al. [4] onderzochten reeds deze Hypercube distributions onder set semantiek en kwamen tot de conclusie dat een conjunctieve query $\mathcal{Q}$ steeds parallel-correct is onder Hypercube distributions voor $\mathcal{Q}$ onder set semantiek. Deze resultaten gelden helaas niet onder bag semantiek, aangezien een Hypercube distribution alle benodigde facts van een valuatie voor $\mathcal{Q}$ mogelijk toekent aan meerdere nodes.

Het aangepaste model op basis van geordende netwerken vermijdt echter dit probleem waarbij een valuatie niet op meerdere nodes gebruikt mag worden onder bag semantiek. Bovendien beïnvloedt het aangepaste model de resultaten onder set semantiek niet, waardoor we kunnen besluiten dat onder het aangepaste model conjunctieve queries parallel-correct zijn onder Hypercube distributions, zowel onder set semantiek als onder bag semantiek.

Contents

Chapter 1

Introduction

Due to the increasing popularity of cloud computing and big data, there is a growing need for data processing in distributed and parallel settings. Although parallel and distributed data management systems have been around for quite some time now, the current demand to execute complex queries in function of large-scale data analytics poses new challenges. Furthermore, distributed environments are nowadays no longer restricted to a couple of servers, but may easily consist of thousands of servers.

In order to facilitate the processing of large amounts of data, Google developed the MapReduce programming model [10]. This model allows developers to take advantage of a distributed environment while abstracting away from some core difficulties related to distributed programming. For example, the MapReduce framework takes care of scheduling the program's execution across the different servers, handles inter-server communication and copes with machine failures. More specifically, A user only needs to define two functions, *map* and *reduce*. The MapReduce environment then parallelizes and executes these functions across the different servers. Hadoop [1], an open-source software framework for storing and processing data in a distributed environment, provides with Hadoop MapReduce an open-source implementation of the MapReduce programming model.

More high-level declarative languages were developed to further facilitate the implementation of programs handling large amounts of data in a distributed environment. Popular languages include Pig [15], developed by Yahoo! and Hive [17], developed by Facebook. Queries in these languages are compiled into MapReduce jobs and executed on Hadoop MapReduce.

More recent systems like Spark [5] combine the MapReduce model with in-memory systems. In contrast to traditional database systems where all the data is managed on a single machine, the complexity of evaluating a query on these modern massively distributed systems is no longer determined by the number of IO requests to external memory. Instead, the complexity of calculating a query over massive datasets distributed over a large amount

of servers is dominated by the necessary amount of communication between the different servers.

Based on these distributed systems, Koutris and Suciu [14] proposed a massively parallel communication model (MPC), based on a cluster of nodes (or servers) using a shared-nothing architecture. In this model, computation is performed in alternating phases of global synchronization and communication between the different servers on the one hand and parallel computation on each server on the other hand. During the latter phase, there is no communication between the different servers, implying that each server performs its computation on the locally available data only.

The MPC model is particularly useful to study the computational complexity of algorithms in a massively distributed environment. Beame, Koutris and Suciu [6, 14] studied the computational complexity of evaluating conjunctive queries in this MPC model.

A special case of queries that can be evaluated in the MPC model are those computable in a single round. These embarrassingly parallel MPC programs are characterized by a distribution phase, during which the data is reshuffled across the different servers according to some distribution policy, followed by a computation phase on each server without further communication between different servers. The final result is obtained by taking the union of the local results on each server. A particular family of distribution policies used to evaluate conjunctive queries in the single-round MPC model are Hypercube distributions [7]. A Hypercube distribution for a conjunctive query $\mathcal{Q}$ distributes the data based on the structure of $\mathcal{Q}$. This technique can be traced back to Ganguly, Silberschatz and Tsur [11] and is studied in the context of MapReduce by Afrati and Ullman [3].

A general framework for reasoning about single-round evaluation algorithms under arbitrary distribution policies is provided by Ameloot et al. [4]. They introduced the following correctness properties for queries and distribution policies:

- *Parallel-correctness:* Given a query $\mathcal{Q}$ and a distribution policy $\boldsymbol{P}$, is it true that the single-round distributed evaluation of $\mathcal{Q}$ according to $\boldsymbol{P}$ will always produce the correct result, independently from the instance of data over which $\mathcal{Q}$ is evaluated?

- *Parallel-correctness transfer:* Given two queries $\mathcal{Q}$ and $\mathcal{Q}'$, is it true that $\mathcal{Q}'$ is parallel-correct under every distribution policy under which $\mathcal{Q}$ is parallel-correct?

The latter is especially useful in a setting of automatic data partitioning for a workload of queries where the aim is to achieve overall optimal performance, as it allows multiple queries to be evaluated without reshuffling the data after each query.

Ameloot et al. [4] then studied these correctness properties for unions of conjunctive queries with inequalities. They described a characterization for both properties, allowing them to provide matching upper and lower bounds for deciding both parallel-correctness and transferability. They proved that testing parallel-correctness is Π_2^p-complete, even for conjunctive queries without unions and inequalities. Deciding transferability on the other hand is Π_3^p-complete.

Geck et al. [12] extended these results by considering parallel-correctness for (unions of) conjunctive queries with negations. They provided a matching upper and lower bound, thereby proving that deciding parallel-correctness for unions of conjunctive queries with negations is coNEXPTIME-complete. This lower bound even holds for conjunctive queries with negations.

The results for the single-round evaluation of conjunctive queries mentioned above focus on set semantics, meaning that possible duplicates in the result are ignored. In practice however, queries are often evaluated under bag semantics, implying that duplicates are not removed from the result, unless explicitly requested. Two reasons for this practical approach are as follows. On the one hand, removing duplicates might be computationally expensive over large datasets. On the other hand, these duplicates might be necessary to correctly perform aggregate functions, like counting or averaging the results. Chaudhuri and Vardi [8] provided a definition for the evaluation of conjunctive queries under bag semantics and studied optimization and containment of conjunctive queries under bag semantics.

In this thesis, we extend the work initiated by Ameloot et al. [4] toward bag semantics. We study the problems of parallel-correctness and transferability for conjunctive queries with inequalities under bag semantics. The former is shown to be equivalent to deciding whether or not every valuation for a given conjunctive query is satisfiable on exactly one node in the network. Based on this characterization, we prove that deciding parallel-correctness is in Π_2^p. This upper bound can be lowered if the class of considered distribution policies is further restricted to deterministic or finite distribution policies. We show that deciding parallel-correctness for this class of distribution policies is coNP-complete by providing a matching upper and lower bound.

The characterization for parallel-correctness indirectly implies that some valuations for a conjunctive query $\mathcal{Q}$ will always be grouped together on the same node, assuming $\mathcal{Q}$ is parallel-correct under the considered distribution policy. We use this observation to provide a characterization for transferability. Based on these results, we show that deciding transferability is in EXPTIME. This upper bound can be improved to Π_2^p if the considered conjunctive queries are limited to conjunctive queries without self-joins.

The provided results under bag semantics are quite different from the

results under set semantics, obtained by Ameloot et al. [4]. We therefore study the relation between set and bag semantics while evaluating conjunctive queries with inequalities. We show that parallel-correctness under bag semantics always implies parallel-correctness under set semantics, but the converse is not necessarily true. If we constrain the queries to strongly minimal conjunctive queries with inequalities and the distribution policies to nonreplicating distribution policies, parallel-correctness under set and bag semantics are equivalent. We also show that for transferability there is no immediate relation between set and bag semantics in general. It follows from our observations on parallel-correctness that transferability under set and bag semantics coincide if the considered queries are restricted to strongly minimal ones and the distribution policies are nonreplicating.

The characterization for parallel-correctness under bag semantics might imply some severe restrictions on possible distribution policies. Depending on the considered query $\mathcal{Q}$, it might even be impossible to construct a distribution policy for $\mathcal{Q}$ effectively using more than one node in the network. We therefore present a slightly different distributed evaluation model. In this modified model, a node only derives a fact according to a valuation if there is no other node in the network that already satisfies this valuation. This modified evaluation model is still executable in a single round, assuming each node has some knowledge about the network and the applied distribution policy.

Under this modified model, we study parallel-correctness and transferability under both set and bag semantics. We show that the characterizations under set semantics remain unchanged under this modified model. Deciding parallel-correctness under bag semantics on the other hand simplifies to testing whether or not each valuation for the considered conjunctive query is satisfiable on at least one node in the network. This simplification results in a different characterization for transferability under bag semantics, allowing us to improve the upper bound for deciding transferability in the general case to Π_2^p. In contrast to the original model, a conjunctive query $\mathcal{Q}$ is always parallel-correct under Hypercube distributions for $\mathcal{Q}$ when considering bag semantics under this modified model.

Outline The necessary definitions and terminology are provided in Chapter 2. In Chapter 3, we describe the notion of parallel-correctness and transferability, as well as a summary of the obtained results under set semantics. We study parallel-correctness and transferability for conjunctive queries with inequalities under bag semantics in Chapter 4. The relation of both parallel-correctness and transferability between set and bag semantics is discussed in Chapter 5. In Chapter 6, we provide a modified distributed evaluation model and study the implications for parallel-correctness and transferability under set and bag semantics. Chapter 7 describes Hypercube distributions

and its relation to parallel-correctness under bag semantics. The obtained results for parallel-correctness and transferability under bag semantics are extended toward unions of conjunctive queries with inequalities in Chapter 8. We conclude in Chapter 9.

Chapter 2

Definitions

In this chapter we introduce the necessary definitions and terminology used throughout this master thesis. The terminology related to (unions of) conjunctive queries and (distributed) evaluations under set semantics is based on the work by Ameloot et al. [4], whereas the terminology related to conjunctive queries under bag semantics is based on the work by Chaudhuri and Vardi [8]. Some adaptations were made to improve the consistency of the terminology between both semantics.

2.1 Queries and instances

We assume an infinite set **dom** of data values that are representable by strings over a fixed alphabet. A *database schema* $\mathcal{D}$ is a finite set of relation names R where every R has arity $ar(R)$. A *fact* $R(d_1, \ldots, d_k)$ is over a database schema $\mathcal{D}$ and a universe $U \subseteq \mathbf{dom}$ where $R \in \mathcal{D}$, $k = ar(R)$ and $d_1, \ldots, d_k \in U$. We use $facts(\mathcal{D}, U)$ to denote the set of all facts over database schema $\mathcal{D}$ and universe $U \subseteq \mathbf{dom}$.

An *annotated fact* f_a is a tuple (f, m) with f a fact and $m \in \mathbb{N}_0$ the multiplicity of f. A *bag of facts* F is a set of annotated facts. Every fact f may appear at most once as an annotated fact in B. That is, $(f, m) \in B$ and $(f', m') \in B$ implies that $f \neq f'$. Intuitively, the multiplicity m of a fact f indicates the number of times f appears in the bag. We denote the set of facts appearing in F by $facts(F)$ and the multiplicity of a fact f in the bag F by $mul(f, F)$. For convenience, we assume $mul(f, F) = 0$ when $f \notin facts(F)$.

When considering two bags of facts F and G, the *bag union* H, denoted $F \cup_B G$, is defined as follows: $facts(H) = facts(F) \cup facts(G)$ and $mul(f, H) = mul(f, F) + mul(f, G)$ for each fact $f \in facts(H)$.

A bag of facts F is a subset of or equal to a bag of facts G, denoted $F \subseteq G$, if for each fact $f \in facts(F)$ it holds that $mul(f, F) \leq mul(f, G)$.

Under set semantics, a *(database) instance* I over $\mathcal{D}$ is a finite set of

facts of $\mathcal{D}$. Under bag semantics, a *(database) instance* I over a database schema $\mathcal{D}$ is a bag of facts, with $facts(I) \subseteq facts(\mathcal{D})$. Under both set and bag semantics, we use $adom(I)$ to denote the set of data values occurring in I.

A *query* $\mathcal{Q}$ *over input schema* $\mathcal{D}_1$ *and output schema* $\mathcal{D}_2$ is a generic mapping from instances over $\mathcal{D}_1$ to instances over $\mathcal{D}_2$.

2.2 Conjunctive queries

Assume an infinite set of variables **var**, disjoint from **dom**. An *atom* over a database schema $\mathcal{D}$ is of the form $R(\mathbf{x})$, with $R \in \mathcal{D}$ and $\mathbf{x} = (x_1, \ldots, x_k)$ a tuple of variables in **var** with $k = ar(R)$.

Conjunctive queries A *conjunctive query* $\mathcal{Q}$ over input schema $\mathcal{D}$ is an expression of the form

$$T(\mathbf{x}) \leftarrow R_1(\mathbf{y_1}), \ldots, R_m(\mathbf{y_m})$$

where every $R_i(\mathbf{y_i})$ is an atom over $\mathcal{D}$ and $T(\mathbf{x})$, the *head atom*, is an atom with $T \notin \mathcal{D}$. Every variable $x \in \mathbf{x}$ needs to appear in at least one $\mathbf{y_i}$. We refer to $T(\mathbf{x})$ as $head_{\mathcal{Q}}$, to the set $\{R_1(\mathbf{y_1}), \ldots, R_m(\mathbf{y_m})\}$ as $body_{\mathcal{Q}}$ and to the set of all variables occurring in $\mathcal{Q}$ as $vars(\mathcal{Q})$. The set of all conjunctive queries is denoted by **CQ**.

A conjunctive query is *without self-joins* if all of its atoms have distinct relation names. A conjunctive query $\mathcal{Q}$ is *full* if every variable occurring in $\mathcal{Q}$ appears in the head atom.

Conjunctive queries with inequalities Conjunctive queries can be extended with inequalities. More formally, a *conjunctive query* $\mathcal{Q}$ *with inequalities* over input schema $\mathcal{D}$ is an expression of the form

$$T(\mathbf{x}) \leftarrow R_1(\mathbf{y_1}), \ldots, R_m(\mathbf{y_m}), \beta_1, \ldots, \beta_p$$

where every β_i is an equality of the form $z \neq z'$. For every such inequality, we require that z and z' are distinct variables occurring in at least one $\mathbf{y_i}$. We refer to this set of inequalities $\{\beta_1, \ldots, \beta_p\}$ as $ineq_{\mathcal{Q}}$. The terminology and constraints mentioned above for conjunctive queries are trivially applicable to conjunctive queries with inequalities as well. Specifically notice that $body_{\mathcal{Q}} = \{R_1(\mathbf{y_1}), \ldots, R_m(\mathbf{y_m})\}$, so the inequalities in $\mathcal{Q}$ are not a part of $body_{\mathcal{Q}}$. We use $\mathbf{CQ}^{\neq}$ to refer to the set of conjunctive queries with inequalities. Notice that $\mathbf{CQ} \subsetneq \mathbf{CQ}^{\neq}$, as a query with no inequalities is just a special case of the broader class of conjunctive queries with inequalities.

A conjunctive query with inequalities is *without self-joins* if all of its atoms have distinct relation names. A conjunctive query with inequalities $\mathcal{Q}$ is *full* if every variable occurring in $\mathcal{Q}$ appears in the head atom.

Unions of conjunctive queries A *union of conjunctive queries* $\mathcal{Q}$ is the union over a finite set of conjunctive queries. More formally, $\mathcal{Q} = \cup_{i=1}^{n} \mathcal{Q}_i$ with $\mathcal{Q}_i \in \mathbf{CQ}^{\neq}$ for every $\mathcal{Q}_i \in \mathcal{Q}$. It is required that every subquery $\mathcal{Q}_i$ uses the same relation name in its head atom. For convenience, we require that no two subqueries of $\mathcal{Q}$ use the same variable. That is, $vars(\mathcal{Q}_i) \cap vars(\mathcal{Q}_j) = \varnothing$ for every $i \neq j$. We use $varmax(\mathcal{Q})$ to denote the maximum number of variables in any subquery $\mathcal{Q}_i$ of $\mathcal{Q}$.

The set of all unions of conjunctive queries with inequalities is denoted by $\mathbf{UCQ}^{\neq}$. Furthermore, we use $\mathbf{UCQ}$ to denote the set of unions of conjunctive queries without inequalities. A union of conjunctive queries $\mathcal{Q}$ is without inequalities if all of its subqueries are without inequalities or, more formally, if $\mathcal{Q}_i \in \mathbf{CQ}$ for every $\mathcal{Q}_i \in \mathcal{Q}$.

2.3 Evaluation of conjunctive queries

A *pre-valuation* for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is a total function $V : vars(\mathcal{Q}) \to \mathbf{dom}$, which naturally extends to atoms and sets of atoms. We say that a pre-valuation V is consistent for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ if for every inequality $z \neq z'$ in $\mathcal{Q}$ it holds that $V(z) \neq V(z')$. A consistent pre-valuation V for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is called a *valuation*. In this case, we refer to $V(body_{\mathcal{Q}})$ as the facts *required* by V. A function V is a valuation for a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ if it is a valuation for a subquery $\mathcal{Q}_i \in \mathcal{Q}$.

When comparing two valuations V_1 and V_2 for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ with $V_1(head_{\mathcal{Q}}) = V_2(head_{\mathcal{Q}})$, we use $V_1 \leq_{\mathcal{Q}} V_2$ to denote the fact that $V_1(body_{\mathcal{Q}}) \subseteq V_2(body_{\mathcal{Q}})$. Analogously, we use $V_1 <_{\mathcal{Q}} V_2$ to denote the fact that $V_1(body_{\mathcal{Q}}) \subsetneq V_2(body_{\mathcal{Q}})$. This notation is also used while comparing valuations for a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$. Let V_1 and V_2 be valuations for respectively $\mathcal{Q}_1$ and $\mathcal{Q}_2$ with $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{Q}$. We write $V_1 \leq_{\mathcal{Q}} V_2$ if $V_1(head_{\mathcal{Q}_1}) = V_2(head_{\mathcal{Q}_2})$ and $V_1(body_{\mathcal{Q}_1}) \subseteq V_2(body_{\mathcal{Q}_2})$. Analogously, we write $V_1 <_{\mathcal{Q}} V_2$ if $V_1(head_{\mathcal{Q}_1}) = V_2(head_{\mathcal{Q}_2})$ and $V_1(body_{\mathcal{Q}_1}) \subsetneq V_2(body_{\mathcal{Q}_2})$.

Valuations under set semantics Under set semantics, a valuation V *satisfies* a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ on an instance I if all facts required by V are in I. In that case, V *derives the fact* $V(head_{\mathcal{Q}})$. We define the *result* of a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ on instance I, denoted $\mathcal{Q}_{\text{set}}(I)$, as the set of facts that can be derived by satisfying valuations for $\mathcal{Q}$ on I. The *result* $\mathcal{Q}_{\text{set}}(I)$ of a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is defined as the set union of the results of all the subqueries of $\mathcal{Q}$ on I.[1]

Definition 2.1. A query $\mathcal{Q}$ is *monotone under set semantics* if for every pair of instances I and I' with $I' \subseteq I$ it holds that $\mathcal{Q}_{\text{set}}(I') \subseteq \mathcal{Q}_{\text{set}}(I)$.

[1] We write $\mathcal{Q}(I)$ instead of $\mathcal{Q}_{\text{set}}(I)$ if it is clear that we are working under set semantics.

As negated atoms aren't allowed, all conjunctive queries $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ under set semantics are trivially monotone.

Valuations under bag semantics Under bag semantics, a valuation V *satisfies* a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ on instance I if $V(body_{\mathcal{Q}}) \subseteq facts(I)$. In that case, V *derives the annotated fact* $f_a = (V(head_{\mathcal{Q}}), m)$, with

$$m = \prod_{a \in body_{\mathcal{Q}}} mul(V(a), I).$$

For convenience, we also say that V derives the fact $f = V(head_{\mathcal{Q}})$ if V satisfies $\mathcal{Q}$ on I. The result of V on an instance I, denoted $[\mathcal{Q}, V]_{\text{bag}}(I)$, is the bag of annotated facts derived by V on instance I. This bag is empty when V doesn't satisfy $\mathcal{Q}$ on I. The *result* $\mathcal{Q}_{\text{bag}}(I)$ of a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ on I is defined as the bag union over all results of satisfying valuations for $\mathcal{Q}$ on I:

$$\mathcal{Q}_{\text{bag}}(I) = \bigcup_{V \in \mathcal{V}}{}_{B} [\mathcal{Q}, V]_{\text{bag}}(I)$$

with $\mathcal{V}$ the set containing all valuations satisfying $\mathcal{Q}$ on I. The *result* $\mathcal{Q}_{\text{bag}}(I)$ of a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is defined as the bag union of the results of all the subqueries of $\mathcal{Q}$ on I.[2]

Definition 2.2. A query $\mathcal{Q}$ is *monotone under bag semantics* if for every pair of instances I and I' with $I' \subseteq I$ it holds that $\mathcal{Q}_{\text{bag}}(I') \subseteq \mathcal{Q}_{\text{bag}}(I)$.

Proposition 2.3. *Conjunctive queries in* $\mathbf{CQ}^{\neq}$ *are monotone under bag semantics.*

Proof. Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a conjunctive query. Let further I and I' be two instances with $I' \subseteq I$. To prove that $\mathcal{Q}_{\text{bag}}(I') \subseteq \mathcal{Q}_{\text{bag}}(I)$, we show the following for each fact $f \in facts(\mathcal{Q}_{\text{bag}}(I'))$: *(i)* $f \in facts(\mathcal{Q}_{\text{bag}}(I))$ and *(ii)* $mul(f, \mathcal{Q}_{\text{bag}}(I')) \leq mul(f, \mathcal{Q}_{\text{bag}}(I))$. To this end, let f be a fact appearing in $facts(\mathcal{Q}_{\text{bag}}(I'))$. Since $f \in facts(\mathcal{Q}_{\text{bag}}(I'))$, there exists a valuation V for $\mathcal{Q}$ with $V(body_{\mathcal{Q}}) \subseteq I'$ and $f = V(head_{\mathcal{Q}})$. As $I' \subseteq I$, it holds that $V(body_{\mathcal{Q}}) \subseteq I$. As a result, V satisfies $\mathcal{Q}$ on I as well, so $f \in facts(\mathcal{Q}_{\text{bag}}(I))$.

Note that the multiplicity $mul(f, \mathcal{Q}_{\text{bag}}(I'))$ of a fact f in the result $\mathcal{Q}_{\text{bag}}(I')$ is by definition the sum of the multiplicities of the annotated facts derived by satisfying valuations V for $\mathcal{Q}$ on I' having $f = V(head_{\mathcal{Q}})$:

$$mul(f, \mathcal{Q}_{\text{bag}}(I')) = \sum_{V \in \mathcal{V}'} mul(f, [\mathcal{Q}, V]_{\text{bag}}(I'))$$

[2]We write $\mathcal{Q}(I)$ and $[\mathcal{Q}, V](I)$ instead of respectively $\mathcal{Q}_{\text{bag}}(I)$ and $[\mathcal{Q}, V]_{\text{bag}}(I)$ if it is clear that we are working under bag semantics.

with $\mathcal{V}'$ the set of valuations V satisfying $\mathcal{Q}$ on I' having $f = V(head_{\mathcal{Q}})$. The calculation of $mul(f, \mathcal{Q}_{\text{bag}}(I))$ is analogue:

$$mul(f, \mathcal{Q}_{\text{bag}}(I)) = \sum_{V \in \mathcal{V}} mul(f, [\mathcal{Q}, V]_{\text{bag}}(I))$$

with $\mathcal{V}$ the set of valuations V satisfying $\mathcal{Q}$ on I having $f = V(head_{\mathcal{Q}})$. So, in order to prove that $mul(f, \mathcal{Q}_{\text{bag}}(I')) \leq mul(f, \mathcal{Q}_{\text{bag}}(I))$, it suffices to show that $\mathcal{V}' \subseteq \mathcal{V}$ and $mul(f, [\mathcal{Q}, V]_{\text{bag}}(I')) \leq mul(f, [\mathcal{Q}, V]_{\text{bag}}(I))$ for each valuation $V \in \mathcal{V}'$.[3] We already proved that each valuation V with $f = V(head_{\mathcal{Q}})$ satisfying $\mathcal{Q}$ on I' also satisfies $\mathcal{Q}$ on I. Thus, $\mathcal{V}' \subseteq \mathcal{V}$.

Observe that, by adding facts to I', the multiplicity of each fact appearing in I' can never go down. As a result, $mul(f, [\mathcal{Q}, V]_{\text{bag}}(I'))$ can only become larger while adding facts to I', because the multiplicity of the fact derived by V is defined as the product of the multiplicities of the facts required for V.[4] Clearly, I can be constructed from I' by adding the missing facts to I', so we conclude that $mul(f, [\mathcal{Q}, V]_{\text{bag}}(I')) \leq mul(f, [\mathcal{Q}, V]_{\text{bag}}(I))$. □

2.4 Networks, data distribution and policies

A *network* $\mathcal{N}$ is a nonempty finite set of values from **dom**, called *nodes*.

A *distribution policy* $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ for a database schema $\mathcal{D}$ and a network $\mathcal{N}$ consists of both a universe U and a total function $rfacts_{\boldsymbol{P}} : \mathcal{N} \to \mathcal{P}(facts(\mathcal{D}, U))$ that maps each node $\kappa \in \mathcal{N}$ onto a set of facts from $facts(\mathcal{D}, U)$. A node $\kappa \in \mathcal{N}$ is responsible for a fact $f \in facts(\mathcal{D}, U)$ under $\boldsymbol{P}$ if $f \in rfacts_{\boldsymbol{P}}(\kappa)$.

Local instances under set semantics Assume a policy $\boldsymbol{P}$ and an instance I for a schema $\mathcal{D}$. The function $loc\text{-}inst_{\text{set},\boldsymbol{P},I}$ maps each node $\kappa \in \mathcal{N}$ onto the set of facts it is responsible for. More formally, this set of facts is defined as $I \cap rfacts_{\boldsymbol{P}}(\kappa)$. In this case, we refer to I as the *global instance* and to $loc\text{-}inst_{\text{set},\boldsymbol{P},I}(\kappa)$ as the *local instance at node* κ.

Local instances under bag semantics Assume a policy $\boldsymbol{P}$ and an instance I for a schema $\mathcal{D}$. The function $loc\text{-}inst_{\text{bag},\boldsymbol{P},I}$ intuitively maps each node $\kappa \in \mathcal{N}$ onto the bag of facts it is responsible for. More formally, $facts(loc\text{-}inst_{\text{bag},\boldsymbol{P},I}(\kappa)) = facts(I) \cap rfacts_{\boldsymbol{P}}(\kappa)$ and for every fact f appearing in $facts(loc\text{-}inst_{\text{bag},\boldsymbol{P},I}(\kappa))$ the multiplicity $mul(f, loc\text{-}inst_{\text{bag},\boldsymbol{P},I}(\kappa))$ equals $mul(f, I)$. Analogously to the terminology used under set semantics,

[3] It is important to note that this condition is only valid when the sum does not contain negative values. But since we are using multiplicities in the sum, negative values are impossible.

[4] Again, we are using the property that the multiplicity is never negative, as a negative factor could actually make the product negative and thus smaller.

we refer to I as the *global instance* and to $\textit{loc-inst}_{\text{bag},\boldsymbol{P},I}(\kappa)$ as the *local instance at node* κ.

Distributed evaluation under set semantics For a distribution policy $\boldsymbol{P}$ over a network $\mathcal{N}$, the result $[\mathcal{Q},\boldsymbol{P}]_{\text{set}}(I)$ of the distributed evaluation under set semantics of a query $\mathcal{Q}$ on an instance I in one round is defined as

$$[\mathcal{Q},\boldsymbol{P}]_{\text{set}}(I) = \bigcup_{\kappa\in\mathcal{N}} \mathcal{Q}_{\text{set}}(\textit{loc-inst}_{\text{set},\boldsymbol{P},I}(\kappa)).$$

Intuitively, $\mathcal{Q}$ is evaluated at each node κ separately, after which the set union of all results is constructed.

Distributed evaluation under bag semantics Given a distribution policy $\boldsymbol{P}$ defined over a network $\mathcal{N}$, the result $[\mathcal{Q},\boldsymbol{P}]_{\text{bag}}(I)$ of the distributed evaluation under bag semantics of a query $\mathcal{Q}$ on an instance I in one round is defined as follows:

$$[\mathcal{Q},\boldsymbol{P}]_{\text{bag}}(I) = \bigcup_{\kappa\in\mathcal{N}}{}_{B} \mathcal{Q}_{\text{bag}}(\textit{loc-inst}_{\text{bag},\boldsymbol{P},I}(\kappa)).$$

This definition closely resembles the definition under set semantics, although we are now working under bag semantics instead of set semantics.

Notice that, by definition of the distribution policy $\boldsymbol{P} = (U, \textit{rfacts}_{\boldsymbol{P}})$, $\textit{mul}(f, I)$ and $\textit{mul}(f, \textit{loc-inst}_{\text{bag},\boldsymbol{P},I}(\kappa))$ are equal for every node κ and every fact f if $f \in \textit{rfacts}_{\boldsymbol{P}}(\kappa)$. In other words, a distribution policy can only decide whether or not a fact is assigned to a node. It cannot change the multiplicity of f on this node. It follows that $\textit{mul}(f, [Q,V]_{\text{bag}}(I)) = \textit{mul}(f, [Q,V]_{\text{bag}}(\textit{loc-inst}_{\text{bag},\boldsymbol{P},I}(\kappa)))$ for each valuation V that is satisfied on the node κ.

2.5 Classes of distribution policies

In order to reason about the complexity of problems with a distribution policy as a part of the input, we need some bound n on the length of strings representing node names and data values. Apart from the classes $\mathcal{P}_{fin}$ and $\mathfrak{P}_{nondet}$, both introduced by Ameloot et al. [4], we describe another class of distribution policies $\mathfrak{P}_{det}$. The latter is closely related to the notion of *PTIME-testable* classes of distribution policies [4].

We first consider a class of distribution policies over finite universes, denoted $\mathcal{P}_{fin}$. A policy $\boldsymbol{P} = (U, \textit{rfacts}_{\boldsymbol{P}})$ belongs to $\mathcal{P}_{fin}$ if it can be specified by an explicit enumeration of the data values in U and an explicit enumeration of all pairs (κ, f) where $f \in \textit{rfacts}_{\boldsymbol{P}}(\kappa)$.

Instead of an explicit enumeration of pairs (κ, f), we could use a "test algorithm" to describe $\textit{rfacts}_{\boldsymbol{P}}$. On an input (κ, f), with κ a node and f a

fact, this "test algorithm" decides whether $f \in \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ with time bound l^k, where $l = |(\kappa, f)|$ is the length of the input and k is a constant. A policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over a network $\mathcal{N}$ is in $\mathcal{P}^k_{nondet}$ if it can be specified by a a pair $(n, \mathcal{A}_P)$, with n a natural number in unary representation and $\mathcal{A}_P$ a non-deterministic algorithm. The value n is used to give an upper bound to the length of data values in the universe U and the names of the nodes in $\mathcal{N}$. More specifically, the universe U consists of all the data values representable by a string of length at most n and the network $\mathcal{N}$ consists of all the nodes representable by strings of length at most n. A fact f is in $\mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ for a given node κ if $\mathcal{A}_P$ has an accepting run of at most $|(\kappa, f)|^k$ steps on input (κ, f). We define $\mathfrak{P}_{nondet}$ as the set $\{\mathcal{P}^k_{nondet} \mid k \geq 2\}$. Note that each policy in $\mathcal{P}_{fin}$ can be described in $\mathcal{P}^2_{nondet}$.

Analogously to $\mathfrak{P}_{nondet}$, we also define a class of distribution policies $\mathfrak{P}_{det}$. A policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ is in $\mathcal{P}^k_{det}$ if it can be specified by a tuple $(\mathcal{N}, n, \mathcal{A}_P)$ where $\mathcal{N}$ is an explicit enumeration of the nodes in the network, n is a natural number in unary representation and $\mathcal{A}_P$ is a deterministic algorithm. The universe U of $\boldsymbol{P}$ is the set of values representable by strings of length at most n. Given a fact f and a node κ, algorithm $\mathcal{A}_P$ decides in at most $|(f, \kappa)|^k$ steps whether $f \in \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. We define $\mathfrak{P}_{det}$ as the set of policies $\{\mathcal{P}^k_{det} \mid k \geq 2\}$. Notice that, in contrast to $\mathfrak{P}_{nondet}$, the description of a policy in $\mathfrak{P}_{det}$ contains an explicit enumeration of the nodes in the network. This explicit enumeration combined with the deterministic test algorithm will prove useful when constructing an improved upper bound for the time complexity of parallel-correctness under bag semantics.

Chapter 3

Literature study

This chapter first introduces the main topics of this master thesis: parallel-correctness and transferability. After that we summarize the results related to unions of conjunctive queries with inequalities under set semantics, obtained by Ameloot et al. [4].

3.1 Parallel-correctness and transferability

Intuitively, the notion of parallel-correctness relates to whether or not the distributed execution of a query with relation to a specific distribution policy produces the correct result. That is, the result should be the same as if the query was executed on the same instance of facts on a single node.

Definition 3.1 ([4])**.** A query $\mathcal{Q}$ is *parallel-correct on instance I under distribution policy $\boldsymbol{P}$* if $\mathcal{Q}(I) = [\mathcal{Q}, \boldsymbol{P}](I)$.

Alternatively, we could define parallel-correctness as a combination of parallel-soundness and parallel-completeness. A query $\mathcal{Q}$ is *parallel-sound* on instance I under distribution policy $\boldsymbol{P}$ if $[\mathcal{Q}, \boldsymbol{P}](I) \subseteq \mathcal{Q}(I)$. Analogously, a query $\mathcal{Q}$ is *parallel-complete* on instance I under distribution policy $\boldsymbol{P}$ if $\mathcal{Q}(I) \subseteq [\mathcal{Q}, \boldsymbol{P}](I)$.

We now lift the previous definition to all instances:

Definition 3.2 ([4])**.** A query $\mathcal{Q}$ is *parallel-correct under distribution policy $\boldsymbol{P}$* if $\mathcal{Q}$ is parallel-correct on all instances I under $\boldsymbol{P}$.

The focus of parallel-correctness is on a single distribution policy $\boldsymbol{P}$ and query $\mathcal{Q}$. If multiple queries need to be evaluated, it might be interesting to study whether or not the property of being parallel-correct is carried over from one query to another. This characteristic might for example be useful in a setting of automatic data partitioning where the aim is to optimize the evaluation of a bunch of queries, as it allows multiple queries to be evaluated

without reshuffling the data after each query evaluation. Informally, parallel-correctness transfer from a query $\mathcal{Q}$ to a query $\mathcal{Q}'$ guarantees that $\mathcal{Q}'$ will be parallel-correct under every distribution policy $\boldsymbol{P}$ under which $\mathcal{Q}$ is parallel-correct.

Definition 3.3 ([4])**.** For two queries $\mathcal{Q}$ and $\mathcal{Q}'$ over the same input schema, *parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$* if $\mathcal{Q}'$ is parallel-correct under every distribution policy for which $\mathcal{Q}$ is parallel-correct. In this case, we write $\mathcal{Q} \xrightarrow{\text{pc}} \mathcal{Q}'$.

Parallel-correctness and transferability are defined over queries in general. In this thesis we focus on parallel-correctness and transferability for conjunctive queries and unions of conjunctive queries.

3.2 Parallel-correctness under set semantics

3.2.1 Characterization

When considering queries in $\mathbf{UCQ}^{\neq}$, the following condition clearly is a sufficient condition for parallel-correctness:

Condition 3.4 ([4])**.** *Let $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ be a query and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over a network $\mathcal{N}$. For every valuation V for $\mathcal{Q}$ over U, there is a node $\kappa \in \mathcal{N}$ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

It is however not a necessary condition for parallel-correctness, as the following example will show:

Example 3.5 ([4])**.** Consider the following conjunctive query $\mathcal{Q}$,

$$T(x, z) \leftarrow R(x, y), R(y, z), R(x, x),$$

and the universe $U = \{a, b\}$. Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over the network $\mathcal{N} = \{\kappa_1, \kappa_2\}$, distributing every fact except $R(a, b)$ onto κ_1 and every fact except $R(b, a)$ onto κ_2.

Notice that Condition 3.4 isn't satisfied, as the required facts for the valuation $V = \{x \mapsto a, y \mapsto b, z \mapsto a\}$ do not meet on at least one of the nodes in $\mathcal{N}$.

The query $\mathcal{Q}$ however is parallel-correct under $\boldsymbol{P}$. First, notice that every fact f derived by a valuation V not requiring both $R(a, b)$ and $R(b, a)$ is trivially satisfied on one of the nodes (or even both nodes, if neither of both facts is required). Therefore, we only need to focus on the valuations requiring both $R(a, b)$ and $R(b, a)$. Clearly, there are only two such valuations: $V_1 = \{x \mapsto a, y \mapsto b, z \mapsto a\}$ and $V_2 = \{x \mapsto b, y \mapsto a, z \mapsto b\}$. The fact $T(a, a)$ derived by V_1 is however derivable by another valuation $V' = \{x \mapsto a, y \mapsto a, z \mapsto a\}$, requiring only $R(a, a)$. Therefore, $T(a, a)$ is derivable on both nodes. The reasoning for the fact $T(b, b)$ derived by V_2 is analogous. We conclude that $\mathcal{Q}$ is indeed parallel-correct under $\boldsymbol{P}$. ■

The previous example illustrates that not every valuation V needs to be satisfied, as long as there is another valuation V' deriving the same fact and requiring only a strict subset of the facts required by V. This observation leads to the definition of minimal valuations.

Definition 3.6 ([4])**.** Let $\mathcal{Q} = \cup_{i=1}^{n} \mathcal{Q}_i$ be a query in $\mathbf{UCQ}^{\neq}$ with subqueries $\mathcal{Q}_1, \ldots, \mathcal{Q}_n \in \mathbf{CQ}^{\neq}$. A valuation V_i for $\mathcal{Q}_i$ is minimal for $\mathcal{Q}$ if there is no valuation W_j for some $\mathcal{Q}_j$ such that $W_j <_{\mathcal{Q}} V_i$.

This definition intuitively states that a valuation V_i is minimal if there is no other valuation W_j deriving the same fact as V_i and requiring only a strict subset of the facts required by V_i. In this context, we say that subquery $\mathcal{Q}_i$ witnesses minimality of V_i for $\mathcal{Q}$.

Based on this notion of minimal valuations, we next present a necessary and sufficient condition for parallel-correctness under set semantics:

Condition 3.7 ([4])**.** *Let $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ be a query and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over a network $\mathcal{N}$. For every minimal valuation V for $\mathcal{Q}$ over U, there is a node $\kappa \in \mathcal{N}$ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proposition 3.8 ([4])**.** *A query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is parallel-correct under distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ if and only if Condition 3.7 is satisfied.*

Let $\boldsymbol{P}$ be a distribution policy and let $\mathcal{Q}$ be a query in $\mathbf{UCQ}^{\neq}$. We say that $\boldsymbol{P}$ *saturates* $\mathcal{Q}$ if Condition 3.7 is satisfied. We say furthermore that $\boldsymbol{P}$ *strongly saturates* $\mathcal{Q}$ if Condition 3.4 is satisfied.

3.2.2 Complexity

We now study the complexity of parallel-correctness under set semantics, focusing on both parallel-correctness on a specific instance (Definition 3.1) as well as the more general form of parallel-correctness lifted to all instances (Definition 3.2). We consider multiple classes of queries and distribution policies. A formal description of both problems is as follows:

	$\mathbf{PCI}(\mathcal{C}, \mathcal{P})$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$,
	instance I
Question:	Is $\mathcal{Q}$ parallel-correct on I under $\boldsymbol{P}$?

	$\mathbf{PC}(\mathcal{C}, \mathcal{P})$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$
Question:	Is $\mathcal{Q}$ parallel-correct under $\boldsymbol{P}$?

In these formal descriptions, $\mathcal{C}$ denotes a query class and $\mathcal{P}$ denotes a class of distribution policies.

Proposition 3.9 ([4]). *Problems* $\mathbf{PCI}(\mathcal{C}, \mathcal{P})$ *and* $\mathbf{PC}(\mathcal{C}, \mathcal{P})$ *are* Π_2^p*-complete for every query class* $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}, \mathbf{UCQ}, \mathbf{UCQ}^{\neq}\}$ *and for every policy class* $\mathcal{P} \in \{\mathcal{P}_{fin} \cup \mathfrak{P}_{nondet}\}$.

We provide the lower bound proof for Proposition 3.9, as it contains a technique that can be used to provide a lower bound for deciding parallel-correctness under bag semantics. The proof is based on a reduction using a well-known Π_2^p-complete problem. We first define this problem before providing the lower bound proof. Given some truth assignment β for a propositional formula ψ, we use $\beta \vDash \psi$ to denote the fact that ψ evaluates to true under β.

	Π_2-QBF
Input:	Formula $\varphi = \forall \mathbf{x} \exists \mathbf{y}\ \psi(\mathbf{x}, \mathbf{y})$ where ψ is a propositional formula in 3-CNF over variables $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$
Question:	Does, for every truth assignment $\beta_{\mathbf{x}}$ on $\mathbf{x}$, exist a truth assignment $\beta_{\mathbf{y}}$ on $\mathbf{y}$ with $(\beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}) \vDash \psi$?

It is well-known that Π_2-QBF is Π_2^p-complete [16].

Proposition 3.10 ([4]). $\mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$ *is* Π_2^p*-hard, even for distribution policies over only two nodes.*

Proof. We construct a polynomial reduction from the problem Π_2-QBF to $\mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$. Since Π_2-QBF is Π_2^p-complete, this construction proves that the problem $\mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$ is Π_2^p-hard.

Let $\varphi = \forall \mathbf{x} \exists \mathbf{y}\ \psi(\mathbf{x}, \mathbf{y})$ be an input for Π_2-QBF over variables $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. We use $C_1, \ldots, C_k$ to denote the disjunctive clauses of ψ with $C_j = (\ell_1^j \vee \ell_2^j \vee \ell_3^j)$ for each j. Each literal ℓ_k^j occurring in a clause C_j represents either a variable z or a negated variable $\neg z$, with $z \in \mathbf{x} \cup \mathbf{y}$.

Based on this propositional formula ψ, we next construct a query $\mathcal{Q} \in \mathbf{CQ}$, a distribution policy $\boldsymbol{P} \in \mathcal{P}_{fin}$ and an instance I serving as the corresponding input for $\mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$.

The query $\mathcal{Q}$ is constructed over the variables w_0, w_1 and $x_i, \overline{x}_i, y_j, \overline{y}_j$ for $i \in \{1, \ldots m\}$ and $j \in \{1, \ldots n\}$. Intuitively, w_0 and w_1 represent the Boolean values true and false, whereas x_i, y_j and $\overline{x}_i, \overline{y}_j$ respectively represent the variables $x_i \in \mathbf{x}$, $y_j \in \mathbf{y}$ and its negation $\neg x_i$,$\neg y_j$. For convenience, we overload the notation of a literal ℓ_k^j as follows: if ℓ_k^j represents a negated variable $\neg z$, then ℓ_k^j denotes the variable $\overline{z}$ as well.

We define $\mathbb{W}$ as the set of all triples over $\{w_0, w_1\}$. Let furthermore $\mathbb{W}^+ = \mathbb{W} \setminus \{(w_0, w_0, w_0)\}$. The construction of the conjunctive query $\mathcal{Q}$ is as follows: $head_{\mathcal{Q}} = H(x_1, \ldots, x_m)$ and $body_{\mathcal{Q}} = \mathcal{A}_{sat} \cup \mathcal{A}_{\psi}$, with

$$\begin{aligned}\mathcal{A}_{sat} = &\{\mathit{True}(w_1), \mathit{False}(w_0), \mathit{Neg}(w_0, w_1), \mathit{Neg}(w_1, w_0)\} \\ &\cup \{C_j(\mathbf{w}) \mid j \in \{1, \ldots, k\}, \mathbf{w} \in \mathbb{W}^+\},\end{aligned}$$

and

$$\begin{aligned}\mathcal{A}_{\psi} = &\{\mathit{Neg}(x_i, \overline{x}_i) \mid i \in \{1, \ldots, m\}\} \cup \{\mathit{Neg}(y_j, \overline{y}_j) \mid j \in \{1, \ldots, n\}\} \\ &\cup \{C_j(\ell_1^j, \ell_2^j, \ell_3^j) \mid j \in \{1, \ldots, k\}\}.\end{aligned}$$

The atoms in $\mathcal{A}_{sat}$ intuitively are consistency atoms, representing valid combinations of opposing values for *Neg*-facts, as well as satisfying combinations of values for C_j-facts. The atoms in $\mathcal{A}_{\psi}$ on the other hand represent the logical structure of ψ by relating each variable with its negation and by relating literals occurring in the same clause with each other.

Analogously to $\mathbb{W}$ and $\mathbb{W}^+$, we define $\mathbb{B}$ as the set of all triples over $\{0, 1\}$ and $\mathbb{B}^+ = \mathbb{B} \setminus \{(0, 0, 0)\}$. Let $U = \{0, 1\}$ be a binary universe. The instance I over U is constructed as follows,

$$\begin{aligned}I = &\{\mathit{True}(1), \mathit{False}(0), \mathit{Neg}(1, 0), \mathit{Neg}(0, 1)\} \\ &\cup \{C_j(\mathbf{b}) \mid j \in \{1, \ldots, k\}, \mathbf{b} \in \mathbb{B}\}.\end{aligned}$$

We define I^- as $\{C_j(0, 0, 0) \mid j \in \{1, \ldots, k\}\}$ and I^+ as $I \setminus I^-$. Let $\mathcal{N} = \{\kappa^+, \kappa^-\}$ be a network over two nodes. The finite distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over $\mathcal{N}$ is constructed as follows: $\mathit{rfacts}_{\boldsymbol{P}}(\kappa^+) = I^+$ and $\mathit{rfacts}_{\boldsymbol{P}}(\kappa^-) = I^-$.

The query $\mathcal{Q}$, instance I and finite policy $\boldsymbol{P}$ are obviously computable in time polynomial in the size of ψ. We next prove that the proposed construction is indeed a reduction. That is, $\langle \mathcal{Q}, \boldsymbol{P}, I \rangle \in \mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$ if and only if $\varphi \in \Pi_2$-QBF.

(if) Assume $\varphi \in \Pi_2$-QBF. We prove that $\langle \mathcal{Q}, \boldsymbol{P}, I \rangle \in \mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$ by showing that each fact f in $\mathcal{Q}(I)$ is in $[\mathcal{Q}, \boldsymbol{P}](I)$ as well. To this end, let $f = H(a_1, \ldots, a_m)$ be an arbitrary fact in $\mathcal{Q}(I)$. We show that f is derived on κ^+, thereby indicating that $f \in [\mathcal{Q}, \boldsymbol{P}](I)$.

Let $\beta_{\mathbf{x}}$ be a truth assignment over all the variables in $\mathbf{x}$ defined by $\beta_{\mathbf{x}}(x_i) = a_i$ for each $i \in \{1, \ldots, m\}$. Since we are working under a binary universe $U = \{0, 1\}$, it can easily be seen that this truth assignment $\beta_{\mathbf{x}}$ is well-defined. By assumption, there is truth assignment $\beta_{\mathbf{y}}$ for the variables in $\mathbf{y}$ such that $(\beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}) \vDash \psi$. We refer to this truth assignment $\beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}$ as β.

Next, consider the valuation V for $\mathcal{Q}$ with $V(w_1) = 1$, $V(w_0) = 0$ and $V(z) = \beta(z)$, $V(\overline{z}) = \beta(\overline{z})$ for every variable $z \in \mathbf{x} \cup \mathbf{y}$. Since $\beta \vDash \psi$,

every clause in ψ is satisfied under β. In other words, $\beta \vDash C_j$ for every $j \in \{1, \ldots, k\}$. Therefore, for every $j \in \{1, \ldots, k\}$ there is a $\mathbf{b} \in \mathbb{B}^+$ such that $V(C_j(\ell_1^j, \ell_2^j, \ell_3^j)) = C_j(\mathbf{b})$. We conclude that all facts in $V(body_{\mathcal{Q}})$ are contained in I^+, implying that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa^+)$. It immediately follows that $f = V(head_{\mathcal{Q}})$ is derived on κ^+.

(only if) The proof is by contraposition. Assume $\varphi \notin \Pi_2$-QBF. We prove that $\langle \mathcal{Q}, \boldsymbol{P}, I \rangle \notin \mathbf{PCI}(\mathbf{CQ}, \mathcal{P}_{fin})$. By assumption, there is a truth assignment $\beta_{\mathbf{x}}$ over the variables in $\mathbf{x}$ such that there is no truth assignment $\beta_{\mathbf{y}}$ over the variables in $\mathbf{y}$ with $(\beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}) \vDash \psi$.

Let $f = H(\beta_{\mathbf{x}}(x_1), \ldots, \beta_{\mathbf{x}}(x_m))$ and let $\beta_{\mathbf{y}}$ be a truth assignment over the variables in $\mathbf{y}$ with $\beta_{\mathbf{y}}(y_j) = 0$ for every variable $y_j \in \mathbf{y}$. Next, consider the truth assignment $\beta = \beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}$. This truth assignment β induces a valuation V for $\mathcal{Q}$ over U as follows: $V(w_1) = 1$, $V(w_0) = 0$ and $V(z) = \beta(z)$, $V(\overline{z}) = \beta(\overline{z})$ for every variable $z \in \mathbf{x} \cup \mathbf{y}$. By construction, this valuation V satisfies $\mathcal{Q}$ on instance I, thus $f \in \mathcal{Q}(I)$.

Notice that it is impossible to derive a fact on κ^-, as $rfacts_{\boldsymbol{P}}(\kappa^-)$ does not contain any *Neg*-facts. On the other hand, f cannot be derived on κ^+ either. Indeed, assume there is a valuation V for $\mathcal{Q}$ over U deriving f on κ^+ and let $\beta_{\mathbf{y}}$ be the truth assignment over variables in $\mathbf{y}$ induced by V. By assumption, $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa^+)$, so V should map all C_j-atoms occurring in $\mathcal{A}_\psi$ onto facts in I^+. But this implies that $(\beta_{\mathbf{x}} \cup \beta_{\mathbf{y}}) \vDash \psi$, contradicting our original assumption.

We conclude that $f \notin [\mathcal{Q}, \boldsymbol{P}](I)$, so $\mathcal{Q}$ is not parallel-correct under distribution policy $\boldsymbol{P}$ on instance I. □

3.3 Transferability under set semantics

3.3.1 Characterization

Based on the notion of minimal valuations, we now introduce query covering:

Definition 3.11 ([4])**.** For two queries $\mathcal{Q} = \cup_{h=1}^{m} \mathcal{Q}_h$ and $\mathcal{Q}' = \cup_{i=1}^{n} \mathcal{Q}'_i$ from $\mathbf{UCQ}^{\neq}$, we say that $\mathcal{Q}$ covers $\mathcal{Q}'$ if for every minimal valuation V' for $\mathcal{Q}'$ (witnessed by $\mathcal{Q}'_i$ for some i), there is a minimal valuation V for $\mathcal{Q}$ (witnessed by $\mathcal{Q}_h$ for some h) such that $V'(body_{\mathcal{Q}'_i}) \subseteq V(body_{\mathcal{Q}_h})$.

This property proves useful for describing a necessary and sufficient condition for parallel-correctness transfer under set semantics:

Proposition 3.12 ([4])**.** *For queries $\mathcal{Q}, \mathcal{Q}' \in \mathbf{UCQ}^{\neq}$, parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if $\mathcal{Q}$ covers $\mathcal{Q}'$.*

Notice that for unions of conjunctive queries minimal valuations are defined over the query itself, not over each subquery separately. In particular,

for a query $\mathcal{Q}$ covering a query $\mathcal{Q}'$, it is not a necessary requirement that each subquery of $\mathcal{Q}'$ is covered by a subquery of $\mathcal{Q}$.

Example 3.13 ([4])**.** To illustrate the fact that a query $\mathcal{Q}$ might cover a query $\mathcal{Q}'$ even if there is a subquery of $\mathcal{Q}'$ not being covered by a subquery of $\mathcal{Q}$, consider the unions of conjunctive queries $\mathcal{Q} = \mathcal{Q}_1$ and $\mathcal{Q}' = \mathcal{Q}_1 \cup \mathcal{Q}_2$, with

$$\mathcal{Q}_1 : H() \leftarrow R(),$$
$$\mathcal{Q}_2 : H() \leftarrow R(), S().$$

It can easily be seen that $\mathcal{Q}_2$ is semantically contained in $\mathcal{Q}_1$, implying that $\mathcal{Q}$ and $\mathcal{Q}'$ are equivalent. It clearly follows that parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. According to Proposition 3.12, $\mathcal{Q}$ therefore covers $\mathcal{Q}'$.

Notice however that not every subquery of $\mathcal{Q}'$ is covered by $\mathcal{Q}$. Indeed, $\mathcal{Q}_2$ is not covered by $\mathcal{Q}_1$, as parallel-correctness doesn't transfer from $\mathcal{Q}_1$ to $\mathcal{Q}_2$. The latter can easily be verified by considering a distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over a network $\mathcal{N} = \{\kappa_R, \kappa_S\}$ with $\mathit{rfacts}_{\boldsymbol{P}}(\kappa_R) = \{R()\}$ and $\mathit{rfacts}_{\boldsymbol{P}}(\kappa_S) = \{S()\}$. ■

3.3.2 Complexity

We next consider the complexity of transferability under set semantics for various query classes $\mathcal{C}$:

	PC-Trans($\mathcal{C}$)
Input:	Queries $\mathcal{Q}, \mathcal{Q}' \in \mathcal{C}$
Question:	Does parallel-correctness transfer from $\mathcal{Q}$ to $\mathcal{Q}'$?

Proposition 3.12 can be used to test whether parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$, although a direct application is not feasible. In order to apply Proposition 3.12 directly, we should check a possibly infinite number of valuations over an infinite domain **dom**. However, Ameloot et al. [4] proved the following proposition:

Proposition 3.14 ([4])**.** *Let $\mathcal{Q} = \cup_{i=1}^{m} \mathcal{Q}_i$ and $\mathcal{Q}' = \cup_{j=1}^{n} \mathcal{Q}'_j$ be two queries in* $\mathbf{UCQ}^{\neq}$ *and let the domain $\boldsymbol{dom}_k = \{1, \ldots, k\}$ be a finite subset of the infinite domain **dom** where $k = max(varmax(\mathcal{Q}), varmax(\mathcal{Q}'))$. The following two conditions are equivalent:*

1. *For every minimal valuation V' for $\mathcal{Q}'$ over **dom** (witnessed by $\mathcal{Q}'_j$ for some j), there is a minimal valuation V for $\mathcal{Q}$ over **dom** (witnessed by $\mathcal{Q}_i$ for some i) such that $V'(body_{\mathcal{Q}'_j}) \subseteq V(body_{\mathcal{Q}_i})$.*

2. *For every minimal valuation V' for $\mathcal{Q}'$ over $\boldsymbol{dom}_k$ (witnessed by $\mathcal{Q}'_j$ for some j), there is a minimal valuation V for $\mathcal{Q}$ over $\boldsymbol{dom}_k$ (witnessed by $\mathcal{Q}_i$ for some i) such that $V'(body_{\mathcal{Q}'_j}) \subseteq V(body_{\mathcal{Q}_i})$.*

In other words, we only need to check Proposition 3.12 for a finite number of valuations over a finite domain $\mathbf{dom}_k$. This observation leads to a bound on the complexity of transferability under set semantics:

Proposition 3.15 ([4]). *Both* **PC-Trans**($\mathbf{UCQ}^{\neq}$) *and* **PC-Trans**($\mathbf{CQ}$) *are* Π_3^p*-complete.*

3.4 Strongly minimal queries

3.4.1 Definition

The notion of minimal valuations leads to a special class of conjunctive queries for which every valuation is minimal.

Definition 3.16 ([4]). A conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is *strongly minimal* if all its valuations are minimal.

This definition of strong minimality extends to conjunctive queries with inequalities and unions of conjunctive queries with inequalities in a natural way.

Definition 3.17 ([4]). A query $\mathcal{Q} = \cup_{j=1}^{n} \mathcal{Q}_j$ in $\mathbf{UCQ}^{\neq}$ is *strongly minimal* if there are no valuations V_i and V_j, witnessed by subqueries $\mathcal{Q}_i, \mathcal{Q}_j \in \mathcal{Q}$, with $V_i <_{\mathcal{Q}} V_j$.

We use $\mathcal{C}[sm]$ to denote the set of all queries in $\mathcal{C}$ which are strongly minimal with $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}, \mathbf{UCQ}, \mathbf{UCQ}^{\neq}\}$.

For queries in $\mathbf{CQ}$, it can easily be seen that full queries and queries without self-joins are always strongly minimal queries. However, this is not a necessary condition.

Example 3.18 ([4]). As a counterexample, consider the following conjunctive query $\mathcal{Q}$,

$$H(x, y) \leftarrow R(x, z), R(y, z), S(z, y).$$

This conjunctive query $\mathcal{Q}$ is strongly minimal, although it is not full and contains a self-join. ■

Ameloot et al. [4] proved that deciding whether a conjunctive query $\mathcal{Q} \in \mathbf{CQ}$ is strongly minimal is coNP-complete. They described a sufficient condition for strong minimality:

Condition 3.19 ([4]). *Let* $\mathcal{Q} \in \mathbf{CQ}$ *be a conjunctive query. Every non-head variable* x *occurs in some* R*-atom at some position* i *and there is no other variable that occurs at position* i *of any* R*-atom.*

Condition 3.19 is not necessary for strongly minimal conjunctive queries:

Example 3.20 ([4])**.** As a counterexample, consider the following conjunctive query $\mathcal{Q}$,

$$H() \leftarrow R(x, y), R(y, x).$$

This conjunctive query $\mathcal{Q}$ does not satisfy Condition 3.19. However, $\mathcal{Q}$ is strongly minimal. Indeed, every valuation V for $\mathcal{Q}$ over a universe U either maps x and y onto the same value a or onto two different values a and b in U. In the former case, the required fact is of the form $R(a, a)$, whereas in the latter case the two required facts are of the form $R(a, b)$ and $R(b, a)$ with $a \neq b$. It immediately follows that there cannot exist two valuations V_1 and V_2 for $\mathcal{Q}$ with $V_1 <_{\mathcal{Q}} V_2$. Therefore, every valuation for $\mathcal{Q}$ is minimal, implying that $\mathcal{Q}$ is strongly minimal. ■

The notion of strongly minimal queries proves useful to lower the complexity of problems related to parallel-correctness. Recall for example Condition 3.4, a sufficient condition for parallel-correctness under set semantics. The only difference between this condition and the necessary and sufficient condition for parallel-correctness under set semantics (Proposition 3.8) is the usage of minimal valuations. Therefore, it can easily be seen that Condition 3.4 becomes a necessary and sufficient condition for parallel-correctness under set semantics. This observation leads to the following improved complexity result:

Proposition 3.21 ([4])**.** *For each class $\mathcal{P}$ of PTIME-testable distribution policies, problems* $\mathbf{PCI}(\mathbf{CQ}[sm], \mathcal{P})$ *and* $\mathbf{PC}(\mathbf{CQ}[sm], \mathcal{P})$ *are in coNP.*

Analogously, the complexity of problems related to transferability is lowered as well for strongly minimal queries. We briefly summarize the results obtained by Ameloot et al. [4]:

Proposition 3.22 ([4])**.**

1. $\mathbf{PC\text{-}Trans}(\mathbf{CQ}[sm], \mathbf{CQ})$ *is NP-complete.*
2. $\mathbf{PC\text{-}Trans}(\mathbf{UCQ}[sm], \mathbf{UCQ})$ *is NP-complete.*
3. $\mathbf{PC\text{-}Trans}(\mathbf{UCQ}^{\neq}[sm], \mathbf{UCQ}^{\neq})$ *is in* Π_2^p.
4. $\mathbf{PC\text{-}Trans}(\mathbf{CQ}^{\neq}[sm], \mathbf{CQ})$ *and* $\mathbf{PC\text{-}Trans}(\mathbf{CQ}[sm], \mathbf{CQ}^{\neq})$ *are* Π_2^p*-hard.*

3.4.2 Relation with minimal queries

The notions of minimal valuations and strongly minimal conjunctive queries are closely related to the classical notion of minimal conjunctive queries [2]. Recall that a conjunctive query $\mathcal{Q} \in \mathbf{CQ}$ is minimal if there is no other conjunctive query $\mathcal{Q}' \in \mathbf{CQ}$ equivalent with $\mathcal{Q}$ having strictly less atoms in its body.

Proposition 3.23 ([4]). *Let $\mathcal{Q}$ be a conjunctive query. For every injective valuation V for $\mathcal{Q}$, the valuation V is minimal if and only if $\mathcal{Q}$ is minimal.*

It can easily be seen that every strongly minimal conjunctive query is a minimal conjunctive query as well. On the other hand, not every minimal conjunctive query is a strongly minimal.

Example 3.24 ([4]). As an example of such a minimal conjunctive query $\mathcal{Q}$ that is not strongly minimal, consider the conjunctive query $\mathcal{Q}$ described in Example 3.5,

$$T(x, z) \leftarrow R(x, y), R(y, z), R(x, x).$$

This conjunctive query $\mathcal{Q}$ is minimal, as we cannot remove one of the atoms in $body_{\mathcal{Q}}$ to obtain an equivalent query with strictly less atoms. Next, consider the valuations $V_1 = \{x \mapsto a, y \mapsto b, z \mapsto a\}$ and $V_2 = \{x \mapsto a, y \mapsto a, z \mapsto a\}$. The required facts for V_1 are $R(a, b)$, $R(b, a)$ and $R(a, a)$, whereas the only required fact for V_2 is $R(a, a)$. Since they both derive the same fact $T(a, a)$, the valuation V_1 is not a minimal valuation for $\mathcal{Q}$. We conclude that $\mathcal{Q}$ is not strongly minimal. ■

Chapter 4

Parallel-correctness and transferability under bag semantics

In this chapter, parallel-correctness and transferability are studied in the context of conjunctive queries with inequalities under bag semantics. For both concepts, we describe a characterization that allows to provide an upper bound on the time complexity of deciding parallel-correctness and transferability for conjunctive queries with inequalities under bag semantics. These upper bounds are further improved by considering possible restrictions on both conjunctive queries and distribution policies.

4.1 Parallel-correctness

Condition 3.4 is a sufficient condition for parallel-correctness under set semantics when evaluating unions of conjunctive queries with inequalities. A slight reformulation of this condition limited to conjunctive queries with inequalities is as follows:

Condition 4.1. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a conjunctive query with inequalities and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over a network $\mathcal{N}$. For every valuation V for $\mathcal{Q}$ over U, there is a node $\kappa \in \mathcal{N}$ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Unfortunately, Condition 4.1 is not sufficient for parallel-correctness under bag semantics.

Example 4.2. For an example showing that Condition 4.1 is not sufficient for parallel-correctness under bag semantics, consider the following conjunctive query $\mathcal{Q}$,

$$T(x, z) \leftarrow R(x, y), R(y, z),$$

and the network $\mathcal{N} = \{\kappa_1, \kappa_2\}$. Let $\boldsymbol{P}$ be a distribution policy over $\mathcal{N}$ that distributes each fact onto every node in $\mathcal{N}$. This distribution policy $\boldsymbol{P}$ clearly satisfies Condition 4.1, as each node is responsible for all facts. Let instance $I = \{(R(a,b),1),(R(b,c),1)\}$. The result $[\mathcal{Q},\boldsymbol{P}](I) = \{(T(a,c),2)\}$ is different from $\mathcal{Q}(I) = \{(T(a,c),1)\}$. We thus conclude that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. ■

As shown by the example, there are unwanted duplicates in the distributed result when the same valuation is satisfied on multiple nodes. We therefore modify Condition 4.1 in the following way:

Condition 4.3. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a conjunctive query with inequalities and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over a network $\mathcal{N}$. For every valuation V for $\mathcal{Q}$ over U, there is exactly one node $\kappa \in \mathcal{N}$ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

In this context, we refer to this particular node κ as the node *responsible* for V.

Condition 4.3 is necessary and sufficient for parallel-correctness. Before proving this proposition, we consider two lemmas first. These two lemmas state that $\mathcal{Q}$ cannot be parallel-correct under a policy $\boldsymbol{P}$ if there is a valuation V for $\mathcal{Q}$ with respectively more than one node or no node in the network on which V is satisfiable.

Lemma 4.4. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is not parallel-correct under $\boldsymbol{P}$ if there exists a valuation V for $\mathcal{Q}$ and more than one node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proof. The proof idea is as follows: A valuation V satisfiable on more than one node would derive the same fact $f = V(head_{\mathcal{Q}})$ multiple times. This would lead to a multiplicity of f that is too high, unless there is some kind of compensation. We show that such a compensation is impossible, thereby proving that $\mathcal{Q}$ cannot be parallel-correct under $\boldsymbol{P}$.

Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. Assume there is a valuation V for $\mathcal{Q}$ and multiple nodes $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Let κ_1 and κ_2 be two such nodes. We prove by contradiction that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. To this end, assume that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. Furthermore, let I be an instance with $facts(I) = V(body_{\mathcal{Q}})$ and $f = V(head_{\mathcal{Q}})$.

Recall that by definition

$$mul(f, \mathcal{Q}(I)) = \sum_{S \in \mathcal{V}} mul(f, [\mathcal{Q}, S](I)) \tag{1}$$

with $\mathcal{V}$ the set of satisfying valuations for $\mathcal{Q}$ on I deriving f. Furthermore,

$$mul(f, [\mathcal{Q}, \boldsymbol{P}](I)) = \sum_{\kappa \in \mathcal{N}} \sum_{T \in \mathcal{V}_\kappa} mul(f, [\mathcal{Q}, T](loc\text{-}inst_{\boldsymbol{P},I}(\kappa))) \tag{2}$$

with $\mathcal{V}_\kappa$ the set of satisfying valuations for $\mathcal{Q}$ on $\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)$ deriving f. Intuitively, there is some relation between the terms in the first equation on the one hand and the terms in the second equation on the other hand. We describe this relation by a function μ mapping terms occurring in the second equation onto terms occurring in the first equation.

Let μ be a function that maps each term $\mathit{mul}(f, [\mathcal{Q}, T](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)))$ in equation 2 onto a term $\mathit{mul}(f, [\mathcal{Q}, S](I))$ in equation 1 in such a way that $T = S$. Observe that μ is a total function. Indeed, a valuation T satisfying $\mathcal{Q}$ on $\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)$ also satisfies $\mathcal{Q}$ on I, since $\mathcal{Q}$ is monotone and $\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa) \subseteq I$. As a result, each term $\mathit{mul}(f, [\mathcal{Q}, T](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)))$ in equation 2 is mapped onto a term $\mathit{mul}(f, [\mathcal{Q}, T](I))$ in equation 1. Recall that $\mathit{mul}(f, [\mathcal{Q}, T](I)) = \mathit{mul}(f, [\mathcal{Q}, T](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)))$, assuming T is a satisfiable valuation for $\mathcal{Q}$ on $\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)$. Thus, μ actually maps terms of equation 2 onto terms of equation 1 with the same value.

Since $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_1)$ and $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$ and since $\mathit{facts}(I) = V(\mathit{body}_\mathcal{Q})$, all facts in I are mapped on both κ_1 and κ_2. In other words, $I = \mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_1) = \mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_2)$. So, V is satisfied on both κ_1 and κ_2. This means that equation 2 contains two terms m_1 and m_2 with m_1 being equal to $\mathit{mul}(f, [\mathcal{Q}, V](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_1)))$ and m_2 being equal to $\mathit{mul}(f, [\mathcal{Q}, V](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_2)))$. The function μ projects these two terms both onto the same term $m = \mathit{mul}(f, [\mathcal{Q}, V](I))$ in equation 1.

Observe that, as μ is a total function and as m_1 and m_2 are mapped onto the same term m, the terms in equation 1 are a strict subset of those in equation 2, unless there is a term m' in equation 1 that is not a part of the image of μ. However, the terms in equation 1 cannot be a subset of equation 2, as this would imply that $\mathit{mul}(f, \mathcal{Q}(I)) < \mathit{mul}(f, [\mathcal{Q}, \boldsymbol{P}](I))$, clearly contradicting our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. We conclude that such a term m' in equation 1 that is not a part of the image of μ must exist.

Let V' be the valuation used in this term m', meaning that $m' = \mathit{mul}(f, [\mathcal{Q}, V'](I))$. By definition of μ, this valuation V' cannot be satisfied on a node κ. Indeed, if V' would be satisfiable on a node κ, then the term $\mathit{mul}(f, [\mathcal{Q}, V'](\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa)))$ would appear in equation 2. But this term would be mapped onto m' by μ, thereby contradicting our assumption that m' is not a part of the image of μ.

Since $m' = \mathit{mul}(f, [\mathcal{Q}, V'](I))$ is a term appearing in equation 1, the valuation V' satisfies $\mathcal{Q}$ on I. We conclude that $V'(\mathit{body}_\mathcal{Q}) \subseteq \mathit{facts}(I)$, and thus $V'(\mathit{body}_\mathcal{Q}) \subseteq \mathit{facts}(\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_1))$.[1] But this implies that V' satisfies $\mathcal{Q}$ on $\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_1)$ as well, contradicting our observation that V' cannot be satisfied on a node κ. We conclude that $\mathcal{Q}$ cannot be parallel-correct under $\boldsymbol{P}$. □

[1] Analogously, We could conclude that $V'(\mathit{body}_\mathcal{Q}) \subseteq \mathit{facts}(\mathit{loc\text{-}inst}_{\boldsymbol{P},I}(\kappa_2))$, but this is not important during the rest of the proof.

Lemma 4.5. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is not parallel-correct under $\boldsymbol{P}$ if there exists a valuation V for $\mathcal{Q}$ with no node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proof. Analogous to the proof idea of Lemma 4.4, this proof idea is based on the observation that the multiplicity of $f = V(head_{\mathcal{Q}})$ will be too low if there is no node responsible for V, unless there is some kind of compensation. During the proof, we show that the only possible compensation is another valuation V' deriving the same fact f on more than one node. According to Lemma 4.4 however, $\mathcal{Q}$ cannot be parallel-correct under $\boldsymbol{P}$ in this case.

The proof is by contradiction. Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$ such that there is a valuation V for $\mathcal{Q}$ over U for which there is no node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Assume $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. Let I be an instance with $facts(I) = V(body_{\mathcal{Q}})$ and let $f = V(head_{\mathcal{Q}})$. Trivially, $f \in facts(\mathcal{Q}(I))$ since V satisfies $\mathcal{Q}$ on I.

Analogous to the proof of Lemma 4.4, we base our reasoning on the equations calculating the multiplicity of f:

$$mul(f, \mathcal{Q}(I)) = \sum_{S \in \mathcal{V}} mul(f, [\mathcal{Q}, S](I)) \tag{1}$$

with $\mathcal{V}$ the set of satisfying valuations for $\mathcal{Q}$ on I deriving f, and

$$mul(f, [\mathcal{Q}, \boldsymbol{P}](I)) = \sum_{\kappa \in \mathcal{N}} \sum_{T \in \mathcal{V}_\kappa} mul(f, [\mathcal{Q}, T](loc\text{-}inst_{\boldsymbol{P},I}(\kappa))) \tag{2}$$

with $\mathcal{V}_\kappa$ the set of satisfying valuations for $\mathcal{Q}$ on $loc\text{-}inst_{\boldsymbol{P},I}(\kappa)$ deriving f. Since $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$, the multiplicity $mul(f, \mathcal{Q}(I))$ of f in $\mathcal{Q}(I)$ needs to be the same as the multiplicity $mul(f, [\mathcal{Q}, \boldsymbol{P}](I))$ of f in $[\mathcal{Q}, \boldsymbol{P}](I)$. In other words, both sums need to give the same result.

We reuse the function μ defined in the proof of Lemma 4.4. Recall from the previous proof that μ is a total function, mapping each term $mul(f, [\mathcal{Q}, T](loc\text{-}inst_{\boldsymbol{P},I}(\kappa)))$ in equation 2 onto a term $mul(f, [\mathcal{Q}, T](I))$ in equation 1. Furthermore remember that, as a result, each term in equation 2 is actually mapped onto a term in equation 1 with the same value.

Since there is no node κ in the network where V is satisfied, there is no term $mul(f, [\mathcal{Q}, V](loc\text{-}inst_{\boldsymbol{P},I}(\kappa)))$ for some node κ in equation 2. So, equation 1 contains a term $mul(f, [Q, V](I))$ that is not a part of the image of μ.

We state that at least two terms in equation 2 are mapped by μ onto the same term in equation 1. Indeed, if μ would map each term in equation 2 onto a different term in equation 1, then the terms in equation 2 would simply be a strict subset of the terms in equation 1, meaning that $mul(f, [\mathcal{Q}, \boldsymbol{P}](I)) < mul(f, \mathcal{Q}(I))$. This would contradict our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$.

Let m_1 and m_2 be two such terms in equation 2 that map onto the same term m in equation 1. By the definition of μ, all three terms m_1, m_2 and m thus use the same valuation V'. This means that m_1 and m_2 are using two different nodes κ_1 and κ_2, because the same valuation V' cannot occur multiple times in equation 2 in terms using the same node. As a result, we see that $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_1)$ and $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$.

However, according to Lemma 4.4, the existence of such a valuation V' that is satisfiable on more than one node poses a contradiction with our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. We conclude that $\mathcal{Q}$ cannot be parallel-correct under $\boldsymbol{P}$. □

Lemma 4.5 closely resembles Condition 3.4, a sufficient condition for parallel-correctness under set semantics. We mention without proof that $\mathcal{Q}_{\text{bag}}(I) \neq [\mathcal{Q}, \boldsymbol{P}]_{\text{bag}}(I)$ if $\mathcal{Q}_{\text{set}}(I) \neq [\mathcal{Q}, \boldsymbol{P}]_{\text{set}}(I)$ for every query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$, distribution policy $\boldsymbol{P}$ and instance I.[2] One might think that these two observations would suffice to prove Lemma 4.5 directly. Unfortunately, this is not the case, as Condition 3.4 is not required for parallel-correctness under set semantics. To this end, assume a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ that is parallel-correct under a distribution policy $\boldsymbol{P}$. According to Proposition 3.8, there might exist a valuation V for $\mathcal{Q}$ for which there is no node in the network responsible for all the required facts for V, as long as this valuation V is not a minimal valuation for $\mathcal{Q}$.

Based on these lemmas, we now prove that Condition 4.3 is necessary and sufficient for parallel-correctness.

Proposition 4.6. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is parallel-correct under $\boldsymbol{P}$ if and only if Condition 4.3 is satisfied.*

Proof. (if) Assume that for every valuation V for $\mathcal{Q}$ there is exactly one node $\kappa \in \mathcal{N}$ that satisfies the facts required by V. Trivially, $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$, as each valuation V that satisfies $\mathcal{Q}$ on a given instance I is also satisfied on exactly one node κ when I is distributed according to $\boldsymbol{P}$.

(only if) The proof is by contraposition. Let V be a valuation for $\mathcal{Q}$ such that there are zero or multiple nodes $\kappa \in \mathcal{N}$ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. It directly follows from Lemma 4.4 and Lemma 4.5 that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. □

[2]The relation of parallel-correctness between set and bag semantics is studied in more detail in Chapter 5. More specifically, Proposition 5.1 says that parallel-correctness under bag semantics implies parallel-correctness under set semantics. The contraposition of this proposition trivially leads to our claim that $\mathcal{Q}_{\text{bag}}(I) \neq [\mathcal{Q}, \boldsymbol{P}]_{\text{bag}}(I)$ if $\mathcal{Q}_{\text{set}}(I) \neq [\mathcal{Q}, \boldsymbol{P}]_{\text{set}}(I)$.

4.2 Parallel-correctness complexity

In this section, we consider the complexity of parallel-correctness for various classes of conjunctive queries and distribution policies. We study the complexity of the problem $\mathbf{PC}(\mathcal{C}, \mathcal{P})$, with $\mathcal{C}$ a class of queries and $\mathcal{P}$ a class of distribution policies.

	$\mathbf{PC}(\mathcal{C}, \mathcal{P})$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$
Question:	Is $\mathcal{Q}$ parallel-correct under $\boldsymbol{P}$?

Proposition 4.7. *The problem $\mathbf{PC}(\mathcal{C}, \mathcal{P})$ is in Π_2^p for every query class $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ and every distribution policy class $\mathcal{P} \in \{\mathcal{P}_{fin}\} \cup \mathfrak{P}_{nondet}$.*

Proof. Let k be fixed and let $\langle \mathcal{Q}, \boldsymbol{P} \rangle$ be an input for $\mathbf{PC}(\mathbf{CQ}^{\neq}, \mathcal{P}^k_{nondet})$, with $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ represented by a tuple $(n, \mathcal{A}_{\boldsymbol{P}})$. According to Proposition 4.6, it suffices to show that there is a Π_2^p-algorithm that checks whether for each valuation V for $\mathcal{Q}$ over U there is exactly one node κ such that $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. This condition can be reformulated as follows:

For every valuation V for Q over U and every pair of nodes κ_1 and κ_2 there is a node κ such that:

- $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ and
- $V(\mathit{body}_{\mathcal{Q}}) \not\subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_1)$ or $V(\mathit{body}_{\mathcal{Q}}) \not\subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$ or $\kappa_1 = \kappa_2$.

Intuitively, the first part of this reformulation states that the facts required for each valuation V should meet at at least one node κ, while the second part ensures that these facts don't meet at more than one node.

Since $f \in \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ can be tested nondeterministically by $\mathcal{A}_{\boldsymbol{P}}$ in time $O(n^k)$ for each of the polynomially many facts $f \in V(\mathit{body}_{\mathcal{Q}})$, deciding whether $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ is in NP. This implies the existence of a verifier M that decides in polynomial time on an input $\langle V(\mathit{body}_{\mathcal{Q}}), \kappa \rangle$ whether $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. In other words, if $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$, there exists a certificate c such that M accepts on input $\langle V(\mathit{body}_{\mathcal{Q}}), \kappa \rangle$. On the other hand, if $V(\mathit{body}_{\mathcal{Q}}) \not\subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$, such a certificate c does not exist.[3]

Using the reformulated conditions and the verifier M, we are able to construct a Π_2^p-algorithm deciding $\mathbf{PC}(\mathbf{CQ}^{\neq}, \mathcal{P}^k_{nondet})$:

[3]Note that we only need to consider certificates with a polynomial length relative to the input. Indeed, the verifier M cannot process longer certificates, as it needs to do so in polynomial time.

For every valuation V for Q over U, for every pair of nodes κ_1 and κ_2 and for every pair of certificates c_1 and c_2, there is a node κ and a certificate c such that:

- The verifier M accepts on input $\langle V(body_{\mathcal{Q}}), \kappa\rangle$ with certificate c and
- The verifier M rejects on input $\langle V(body_{\mathcal{Q}}), \kappa_1\rangle$ with certificate c_1 or M rejects on input $\langle V(body_{\mathcal{Q}}), \kappa_2\rangle$ with certificate c_2 or $\kappa_1 = \kappa_2$.

Note that this result also holds for query class **CQ** and policy class $\mathcal{P}_{fin}$, since $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$ and $\mathcal{P}_{fin} \subseteq \mathcal{P}^2_{nondet}$. □

Observe that the resulting upper bound given by Proposition 4.7 is partially due to the existential quantifier over the nodes in the network $\mathcal{N}$ and the certificates for the verifier M. Therefore, we could improve the upper bound if we could drop this existential quantifier. A possible approach is to limit ourselves to networks containing only a polynomial number of nodes on the one hand and deterministic algorithms to decide whether $f \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ for a given fact f and node κ on the other hand. The distribution classes in $\mathfrak{P}_{det}$ satisfy these conditions, resulting in the following upper bound:

Proposition 4.8. *The problem* $\mathbf{PC}(\mathcal{C}, \mathcal{P})$ *is in* coNP *for every query class* $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ *and every distribution policy class* $\mathcal{P} \in \mathfrak{P}_{det} \cup \{\mathcal{P}_{fin}\}$.

Proof. It suffices to show that the complement of the problem $\mathbf{PC}(\mathcal{C}, \mathcal{P})$, denoted $\overline{\mathbf{PC}(\mathcal{C}, \mathcal{P})}$, is in NP for every query class $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ and every distribution policy class $\mathcal{P} \in \mathfrak{P}_{det} \cup \{\mathcal{P}_{fin}\}$.

	$\overline{\mathbf{PC}(\mathcal{C}, \mathcal{P})}$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$
Question:	Is $\mathcal{Q}$ not parallel-correct under $\boldsymbol{P}$?

Let k be fixed. We construct a nondeterministic algorithm that decides $\overline{\mathbf{PC}(\mathbf{CQ}^{\neq}, \mathcal{P}^{\mathbf{k}}_{\mathbf{det}})}$ on input $\langle \mathcal{Q}, \boldsymbol{P}\rangle$ with $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and where $\boldsymbol{P}$ is represented by a tuple $(\mathcal{N}, n, \mathcal{A}_P)$. According to Proposition 4.6, $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$ if and only if there is a valuation V for $\mathcal{Q}$ over U such that the total number of nodes κ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ is zero or at least two. We use this property to construct the nondeterministic algorithm as follows: Guess a valuation V and count the number of nodes κ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. The algorithm rejects if this count equals one. Otherwise, it accepts.

Since valuations are mappings of variables appearing in $\mathcal{Q}$ to values in U and since values in U can be represented by a string of length n or less, it is possible to guess a valuation V in polynomial time. Furthermore, the number of nodes κ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ can also be determined in

polynomial time, as the nodes are an explicit part of the input and $\mathcal{A}_P$ can be used to decide $V(body_Q) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ in polynomial time for each node κ. Thus, The nondeterministic algorithm described above runs in polynomial time.

Notice that this reasoning holds for conjunctive queries in **CQ** as well, since $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$. □

We conclude this section by providing a lower bound for the problem $\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})$. This lower bound immediately infers a lower bound on every problem $\mathbf{PC}(\mathcal{C}, \mathcal{P})$ with query class $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ and distribution policy class $\mathcal{P} \in \{\mathcal{P}_{fin}\} \cup \mathfrak{P}_{nondet} \cup \mathfrak{P}_{det}$. The construction of this lower bound proof resembles the construction used in the proof of Proposition 3.10.

Before describing this lower bound proof, we first briefly mention the problem 3-SAT, as it is used in a reduction during our proof.

	3-SAT
Input:	Propositional formula ψ in 3-CNF over variables $\mathbf{x} = (x_1, \ldots, x_n)$
Question:	Does a truth assignment $\beta_{\mathbf{x}}$ on $\mathbf{x}$ exist with $\beta_{\mathbf{x}} \vDash \psi$?

It is well-known that 3-SAT is NP-complete [9, 13].

Proposition 4.9. *The problem* $\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})$ *is* coNP*-complete, even over networks with only two nodes.*

Proof. Proposition 4.8 already states that $\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})$ is in coNP, so we only need to prove that $\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})$ is coNP-hard.

We construct a polynomial reduction from 3-SAT to $\overline{\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})}$. Since 3-SAT is NP-complete, this construction proves that $\overline{\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})}$ is NP-hard. It immediately follows that $\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})$ is coNP-hard.

Let ψ be an input for 3-SAT over variables $\mathbf{x} = (x_1, \ldots, x_n)$. We use $C_1, \ldots, C_k$ to denote the disjunctive clauses of ψ with $C_j = (\ell_1^j \vee \ell_2^j \vee \ell_3^j)$ for each j. Each literal ℓ_k^j occurring in a clause C_j represents either a variable x_i or a negated variable $\neg x_i$, with $x_i \in \mathbf{x}$.

Based on this propositional formula ψ, we next construct a query $\mathcal{Q} \in \mathbf{CQ}$ and distribution policy $\boldsymbol{P} \in \mathcal{P}_{fin}$ serving as the corresponding input for $\overline{\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})}$.

The query $\mathcal{Q}$ is constructed over the variables $x_i, \overline{x}_i$ for $i \in \{1, \ldots n\}$. Intuitively, x_i and $\overline{x}_i$ respectively represent the variable $x_i \in \mathbf{x}$ and its negation $\neg x_i$. For convenience, we overload the notation of a literal ℓ_k^j as follows: if ℓ_k^j represents a negated variable $\neg x_i$, then ℓ_k^j denotes the variable $\overline{x}_i$ as well.

The construction of the conjunctive query $\mathcal{Q}$ is as follows: $\mathit{head}_\mathcal{Q} = H()$ and

$$\mathit{body}_\mathcal{Q} = \{\mathit{Neg}(x_i, \overline{x}_i) \mid i \in \{1, \ldots, n\}\} \cup \{C_j(\ell_1^j, \ell_2^j, \ell_3^j) \mid j \in \{1, \ldots, k\}\}.$$

Unlike the proof of Proposition 3.10, the variables used in the head atom are not important during the rest of this proof. Therefore, we do not include any variables in $\mathit{head}_\mathcal{Q}$. The atoms in $\mathit{body}_\mathcal{Q}$ intuitively represent the logical structure of ψ by relating each variable with its negation and by relating literals occurring in the same clause with each other.

We define $\mathbb{B}$ as the set of all triples over $\{0, 1\}$ and $\mathbb{B}^+ = \mathbb{B} \setminus \{(0, 0, 0)\}$. Let $\mathcal{N} = \{\kappa_1, \kappa_2\}$ be a network over two nodes and let $U = \{0, 1\}$ be a binary universe. The finite distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over $\mathcal{N}$ is constructed as follows: each fact over *Neg* and each fact over a C_j is assigned to κ_1 and

$$\mathit{rfacts}_{\boldsymbol{P}}(\kappa_2) = \{\mathit{Neg}(0, 1), \mathit{Neg}(1, 0)\} \cup \{C_j(\mathbf{b}) \mid j \in \{1, \ldots, k\}, \mathbf{b} \in \mathbb{B}^+\}.$$

The intuition behind this construction is that the required facts for every valuation V for $\mathcal{Q}$ over U are assigned to κ_1, whereas κ_2 only contains all the required facts for valuations that can be related to truth assignments on $\mathbf{x}$ satisfying ψ.

The query $\mathcal{Q}$ and finite policy $\boldsymbol{P}$ are obviously computable in time polynomial in the size of ψ. We next prove that $\psi \in$ 3-SAT if and only if $\langle \mathcal{Q}, \boldsymbol{P} \rangle \in \overline{\mathbf{PC}(\mathbf{CQ}, \mathcal{P}_{fin})}$.

(if) The proof is by contraposition. Assuming $\psi \notin$ 3-SAT, we show that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. According to Proposition 4.6, this is the case if for every valuation V for $\mathcal{Q}$ there exists exactly one node $\kappa \in \mathcal{N}$ with $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. For every valuation V for $\mathcal{Q}$ it clearly holds that $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_1)$. We therefore only need to show that there is no valuation V for $\mathcal{Q}$ with $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$.

To this end, assume that there is a valuation V for $\mathcal{Q}$ over U with $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$. Now consider the truth assignment $\beta_\mathbf{x}$ on $\mathbf{x}$ having $\beta_\mathbf{x}(x_i) = V(x_i)$ for every variable $x_i \in \mathbf{x}$. It can easily be seen that this truth assignments is well-defined since U is a binary universe, thereby implying that V maps each variable x_i onto either 0 or 1.

Since the only facts over *Neg* are $\mathit{Neg}(0, 1)$ and $\mathit{Neg}(1, 0)$, and since $\mathit{body}_\mathcal{Q}$ contains the atom $\mathrm{Neg}(x_i, \overline{x}_i)$ for every variable x_i, we conclude that $V(x_i) = 1$ if and only if $V(\overline{x}_i) = 0$, and vice versa. It immediately follows that $\beta_\mathbf{x}(\neg x_i)$ equals $V(\overline{x}_i)$. In other words, this valuation V correctly assigns the negated value of a variable x_i to the variable $\overline{x}_i$ representing the negated literal $\neg x_i$.

We now show that this truth assignment $\beta_\mathbf{x}$ satisfies ψ. To this end, consider an arbitrary clause $C_j = (\ell_1^j \vee \ell_2^j \vee \ell_3^j)$ in ψ. Since $C_j(\ell_1^j, \ell_2^j, \ell_3^j) \in \mathit{body}_\mathcal{Q}$

and since $C_j(0,0,0) \notin \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$, the valuation V should map at least one literal ℓ_k^j to 1. By construction of $\beta_{\mathbf{x}}$, it directly follows that $\beta_{\mathbf{x}}(\ell_k^j) = 1$. Thus, this clause C_j is satisfied under $\beta_{\mathbf{x}}$. We conclude that $\beta_{\mathbf{x}}$ indeed satisfies ψ, as this reasoning is clearly applicable to every clause C_j appearing in ψ.

However, this truth assignment $\beta_{\mathbf{x}}$ satisfying ψ poses a contradiction because we initially assumed that ψ is not satisfiable. Therefore, we conclude that there is no valuation V for $\mathcal{Q}$ over U with $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$.

(only if) Assume $\psi \in$ 3-SAT. We prove that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$ by providing a valuation V for $\mathcal{Q}$ over U that is satisfiable on both nodes in $\mathcal{N}$.

Let $\beta_{\mathbf{x}}$ be truth assignment over $\mathbf{x}$ satisfying ψ (this truth assignment exists by assumption). Based on this truth assignment $\beta_{\mathbf{x}}$, we define the valuation V for $\mathcal{Q}$ over U as follows: $V(x_i) = \beta_{\mathbf{x}}(x_i)$ for every variable x_i and $V(\overline{x}_i) = \beta_{\mathbf{x}}(\neg x_i)$ for every variable $\overline{x}_i$.

We next show that $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$. The first part of $\mathit{body}_{\mathcal{Q}}$ requires the fact $V(\mathit{Neg}(x_i, \overline{x}_i))$ for each variable x_i and its negated variable $\overline{x}_i$. By definition of V, it can easily be seen that $V(x_i) = 1$ if and only if $V(\overline{x}_i) = 0$, and vice versa. The only possible required facts for V therefore are $\mathit{Neg}(0,1)$ and $\mathit{Neg}(1,0)$, and these facts are indeed mapped onto κ_2.

The second part of $\mathit{body}_{\mathcal{Q}}$ requires the fact $V(C_j(\ell_1^j, \ell_2^j, \ell_3^j))$ for every clause C_j. Since $\beta_{\mathbf{x}}$ satisfies ψ, we know that there is at least one literal ℓ_k^j appearing in C_j with $\beta_{\mathbf{x}}(\ell_k^j) = 1$. By construction of V, it follows that $V(\ell_k^j) = 1$. In other words, at least one of the three values appearing in the fact $V(C_j(\ell_1^j, \ell_2^j, \ell_3^j))$ has to be 1. We conclude that $V(C_j(\ell_1^j, \ell_2^j, \ell_3^j)) \in \{C_j(\mathbf{b}) \mid \mathbf{b} \in \mathbb{B}^+\}$, and the node κ_2 is by construction indeed responsible for this fact.

We conclude that all the required facts for this valuation V are available on κ_2. These required facts are obviously available on κ_1 as well because every fact is by construction of $\boldsymbol{P}$ assigned to this node κ_1. Since V is satisfiable on both nodes, it immediately follows from Proposition 4.6 that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. □

4.3 Parallel-correctness transfer

4.3.1 Conditions for parallel-correctness transfer

According to proposition 4.6, a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is parallel-correct under a distribution policy $\boldsymbol{P}$ if and only if the required facts for each valuation V meet at exactly one node. We use this property to formulate a sufficient condition for parallel-correctness transferring from $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ to $\mathcal{Q}' \in \mathbf{CQ}^{\neq}$:

Condition 4.10. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. For each valuation V' for $\mathcal{Q}'$ over a universe U, there is a valuation V for $\mathcal{Q}$ over U such that $V'(body_{\mathcal{Q}'}) = V(body_{\mathcal{Q}})$.*

Proposition 4.11. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. If Condition 4.10 satisfies then parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$.*

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$ and assume Condition 4.10 holds. We prove that parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. To this end, assume a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over some universe U under which $\mathcal{Q}$ is parallel-correct. We need to show that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well.

Let V' be a valuation for $\mathcal{Q}'$ over U. According to Condition 4.10, there is a valuation V for $\mathcal{Q}$ over U such that $V'(body_{\mathcal{Q}'}) = V(body_{\mathcal{Q}})$. Since $\mathcal{Q}$ is parallel-correct, there is exactly one node κ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ as stated by Proposition 4.6. Furthermore, since $V'(body_{\mathcal{Q}'}) = V(body_{\mathcal{Q}})$, it also holds that κ is the only node such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. This argumentation holds for each valuation V' for $\mathcal{Q}'$ over U. In other words, for each valuation V' for $\mathcal{Q}'$ over U, there is exactly one node κ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. By Proposition 4.6, we conclude that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well. □

Note that it is not sufficient to require $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$ instead of $V'(body_{\mathcal{Q}'}) = V(body_{\mathcal{Q}})$, as this would only guarantee that the facts in $V'(body_{\mathcal{Q}'})$ would meet at *at least* one node. For parallel-correctness under bag-semantics we require however that these facts meet at *exactly* one node.

Although Condition 4.10 is sufficient for parallel-correctness transfer, it is not a necessary condition, as the following example shows:

Example 4.12. Consider the following two conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$:

$$\mathcal{Q} : H(x, y) \leftarrow R(x, x), R(x, y), R(y, x).$$
$$\mathcal{Q}' : H(x, y) \leftarrow R(x, x), R(x, y).$$

Let $V' = \{x \mapsto a, y \mapsto b\}$ be a valuation for $\mathcal{Q}'$. This valuation requires the following facts $V'(body_{\mathcal{Q}'}) = \{R(a, a), R(a, b)\}$. Since there is no valuation V for $\mathcal{Q}$ with $V(body_{\mathcal{Q}}) = \{R(a, a), R(a, b)\}$, Condition 4.10 is not satisfied.

However, parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. We show that $\mathcal{Q}'$ is parallel-correct under every distribution policy for which $\mathcal{Q}$ is parallel-correct. To this end, let $\boldsymbol{P}$ be such a policy for which $\mathcal{Q}$ is parallel-correct. We need to prove that the required facts $V'(body_{\mathcal{Q}'})$ for each valuation V' over $\mathcal{Q}'$ meet at exactly one node κ under $\boldsymbol{P}$. Let $V' = \{x \mapsto a, y \mapsto b\}$ be such a valuation for $\mathcal{Q}'$. Consider the valuations $V_1 = \{x \mapsto a, y \mapsto a\}$ and $V_2 = \{x \mapsto a, y \mapsto b\}$ for $\mathcal{Q}$. Since $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$, The facts in $V_1(body_{\mathcal{Q}}) = \{R(a, a)\}$ and $V_2(body_{\mathcal{Q}}) = \{R(a, a), R(a, b), R(b, a)\}$ meet at

exactly one node κ.[4] As $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$, we conclude that the facts in $V'(body_{\mathcal{Q}'})$ also meet at only this node κ. Observe that this argumentation still holds when both variables x and y are mapped onto the same value a. In this case, $V'(body_{\mathcal{Q}'})$ would furthermore equal $V_1(body_{\mathcal{Q}})$, which also trivially leads to the conclusion that the facts in $V'(body_{\mathcal{Q}'})$ meet at exactly one node κ. ■

In the previous example, it is shown that Condition 4.10 is too strict to be a necessary condition. We now propose a more general condition for parallel-correctness transfer based on Example 4.12:

Condition 4.13. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$. For each valuation V' for $\mathcal{Q}'$ over some universe U, there exist two valuations V_1 and V_2 for $\mathcal{Q}$ over U such that $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$ and $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$.*

Proposition 4.14. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. If Condition 4.13 satisfies then parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$.*

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$. Assume for each valuation V' for $\mathcal{Q}'$ over some universe U there exist two valuations V_1 and V_2 for $\mathcal{Q}$ over U such that $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$ and $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$. We prove that parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. To this end, Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a universe U under which $\mathcal{Q}$ is parallel-correct. We need to show that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well.

Let V' be a valuation for $\mathcal{Q}'$ over U and let V_1 and V_2 be the two valuations for $\mathcal{Q}$ over U such that $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$ and $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$. By Proposition 4.6, the required facts for both V_1 and V_2 meet at exactly one node. Furthermore, since both $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$ and $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$, it trivially follows that $V_1(body_{\mathcal{Q}}) \subseteq V_2(body_{\mathcal{Q}})$. This implies that the node κ responsible for V_1 is the same as the node responsible for V_2. Since $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$, the required facts for V' meet at this node κ.

Also note that there cannot be another node κ' different from κ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$. Indeed, if such a node κ' would exist, it would be responsible for the required facts for V_1 as well, as $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. This would imply that the required facts for V_1 meet at more than one node, contradicting our initial assumption that $\mathcal{Q}$ is parallel-correct.

Since this argumentation holds for each valuation V' for $\mathcal{Q}'$ over U, we conclude that for each valuation V' for $\mathcal{Q}'$ there is exactly one node κ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Thus, according to Proposition 4.6, $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well. □

[4]Notice that, since $V_1(body_{\mathcal{Q}}) \subseteq V_2(body_{\mathcal{Q}})$, the facts in $V_1(body_{\mathcal{Q}})$ cannot meet at another node than the node where the facts in $V_2(body_{\mathcal{Q}})$ meet. Otherwise, the facts in $V_1(body_{\mathcal{Q}})$ would meet at multiple nodes.

Condition 4.10 clearly is a special case of Condition 4.13, as assuming $V_1 = V_2 = V$ in Condition 4.13 trivially leads to Condition 4.10. However, Condition 4.13 is still not a necessary condition for parallel-correctness transfer, as the following example will show:

Example 4.15. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be the following conjunctive queries:

$$\mathcal{Q} : H(x, z) \leftarrow R(x, y), R(y, z).$$
$$\mathcal{Q}' : H(w, z) \leftarrow R(w, x), R(x, y), R(y, z).$$

Let $V' = \{w \mapsto a, x \mapsto b, y \mapsto c, z \mapsto d\}$ be a valuation for $\mathcal{Q}'$. Trivially there is no valuation V_2 for $\mathcal{Q}$ such that $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$, as $V_2(body_{\mathcal{Q}})$ can contain at most two facts, while $V'(body_{\mathcal{Q}'})$ contains three facts.

We'll show that parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$, thereby showing that Condition 4.13 is not a necessary condition for transferability. To this end, assume a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ under which $\mathcal{Q}$ is parallel-correct. We'll prove that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well.

First, note that for each valuation V' for $\mathcal{Q}'$ there clearly is a valuation V_1 for $\mathcal{Q}$ such that $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. As stated earlier in the proof of Proposition 4.14, this implies that the required facts for V' can never meet at multiple nodes. Thus, it further suffices to show that the required facts for each valuation V' for $\mathcal{Q}'$ meet at *at least one* node under $\boldsymbol{P}$ in order to conclude that the required facts for each valuation V' for $\mathcal{Q}'$ meet at exactly one node, thereby implying that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$.

We now consider how $\boldsymbol{P}$ might distribute facts over the different nodes, assuming we're working under a random universe U. Let $a \in U$ be a random value. Clearly, the fact $R(a, a)$ needs to be mapped onto exactly one node κ, since $H(a, a)$ can be derived by $\mathcal{Q}$ when using only $R(a, a)$. Mapping $R(a, a)$ on multiple nodes or no node at all would therefore make $\mathcal{Q}$ no longer parallel-correct under $\boldsymbol{P}$.

Let now $b \in U$ be another random value. Trivially, there exists a valuation for $\mathcal{Q}$ requiring the facts $R(a, a)$ and $R(a, b)$ and another valuation for $\mathcal{Q}$ requiring $R(a, a)$ and $R(b, a)$. Since $R(a, a)$ is only assigned to κ, we'll need to assign both $R(a, b)$ and $R(b, a)$ to κ as well. In other words, assigning the fact $R(a, a)$ onto the node κ implies that all other facts using the value a will be assigned to κ as well.

Now consider the fact $R(b, b)$. We can trivially apply the argumentation for $R(a, a)$ to $R(b, b)$ as well. Thus, we know that $R(b, b)$ is assigned to exactly one node κ', and that all facts using the value b are assigned to this node κ' as well.

Note that the facts $R(a, b)$ and $R(b, a)$ are assigned to both κ and κ'. Furthermore, the valuation $W = \{x \mapsto a, y \mapsto b, z \mapsto a\}$ for $\mathcal{Q}$ would require exactly these two facts. We conclude that $\kappa = \kappa'$, since the existence of a valuation W that is satisfiable onto more than one node would contradict our

assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. This argumentation holds for each pair of values a and b in the universe U, and thus we conclude that all facts of the form $R(a, a)$, with $a \in U$, are assigned onto *the same* node κ. We already argued that a fact of the form $R(a, a)$ mapped onto a node κ implies that all other facts using a are mapped onto this node κ as well. Thus, each fact using a value from the universe U is mapped onto this node κ. Since each fact effectively uses two values from the universe U, we conclude that *all* facts are assigned to the same node κ.[5]

Since $\mathcal{Q}$ requires $\boldsymbol{P}$ to assign all facts onto a node κ, it trivially follows that the required facts for each valuation V' for $\mathcal{Q}'$ meet at at least one node under $\boldsymbol{P}$. ■

The previous example shows that Condition 4.13 is not a necessary condition for parallel-correctness transfer by constructing a counterexample having a valuation V' for $\mathcal{Q}'$ such that there is no valuation V_2 for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$. The other part of Condition 4.13, requiring a valuation V_1 for $\mathcal{Q}$ for each valuation V' for $\mathcal{Q}'$ such that $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$, is still satisfied in the example. One might ask if it would also be possible to construct a counterexample having a valuation V' for $\mathcal{Q}'$ such that there is no valuation V_1 for $\mathcal{Q}$ with $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. The answer turns out to be no, as we will prove next.

Condition 4.16. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$. For each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$.*

Proposition 4.17. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$. Condition 4.16 is a necessary condition for parallel-correctness transferring from $\mathcal{Q}$ to $\mathcal{Q}'$.*

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$. Assume Condition 4.16 does not hold. That is, there exists a valuation V' for $\mathcal{Q}'$ over a universe U such that there is no valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. We show that parallel-correctness cannot transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ by constructing a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ with two nodes κ_1 and κ_2 under which $\mathcal{Q}$ is parallel-correct, but $\mathcal{Q}'$ is not.

This distribution policy $\boldsymbol{P}$ maps each fact f over U onto κ_1. Furthermore, it also maps each fact $f \in V'(body_{\mathcal{Q}'})$ onto κ_2. Since $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_1)$ and $V'(body_{\mathcal{Q}'}) = rfacts_{\boldsymbol{P}}(\kappa_2)$, valuation V' can be satisfied on both nodes. According to Proposition 4.6, $\mathcal{Q}'$ is not parallel-correct under $\boldsymbol{P}$.

[5]Notice that this doesn't imply that all other nodes have no facts assigned to them. For example, the fact $R(a, b)$ could be assigned to a node κ' different from κ as well. However, these extra nodes will always produce the empty result when $\mathcal{Q}$ is evaluated over them. Indeed, a nonempty result would imply the existence of a valuation that is satisfiable on both κ and κ', contradicting the assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$.

All valuations V for $\mathcal{Q}$ can trivially be satisfied on κ_1. A valuation V for $\mathcal{Q}$ can however never be satisfied on κ_2, as this requires $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$. This requirement cannot be satisfied, as it would trivially imply $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$, thereby contradicting our assumption that there is no valuation V for $\mathcal{Q}$ with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. We conclude that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. □

Let $\mathcal{Q}$ and $\mathcal{Q}'$ be two queries in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy under which $\mathcal{Q}$ is parallel-correct. If Condition 4.16 is satisfied, we know for sure that there is at most one node κ for every valuation V' for $\mathcal{Q}'$ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Indeed, if two or more such nodes would exist, there would be a valuation V for $\mathcal{Q}$ that would be satisfiable on these nodes as well, thereby contradicting our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$.

Condition 4.16 is however not a sufficient condition for transferability, as it only enforces that there is at most one node κ for every valuation V' for $\mathcal{Q}'$ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, while parallel-correctness requires that every valuation V' for $\mathcal{Q}'$ is satisfiable on exactly one node. In other words, we still need to make sure that every valuation V' for $\mathcal{Q}'$ is satisfiable on at least one node κ. Condition 4.13 enforces this requirement by demanding a valuation V_2 for $\mathcal{Q}$ for each valuation V' for $\mathcal{Q}'$ such that $V'(body_{\mathcal{Q}'}) \subseteq V_2(body_{\mathcal{Q}})$. Example 4.15 however shows that Condition 4.13 is not a necessary condition. This example indicates that some valuations for a conjunctive query $\mathcal{Q}$ are guaranteed to be satisfiable on the same node under every distribution policy $\boldsymbol{P}$ under which $\mathcal{Q}$ is parallel-correct.

4.3.2 The set *impFacts*

In the previous section we observed that, depending on the conjunctive query $\mathcal{Q}$, the characterization for parallel-correctness described in Proposition 4.6 implicitly requires that certain valuations for $\mathcal{Q}$ are grouped onto the same node. In other words, there always is a single node in the network responsible for all these valuations. It immediately follows that for every valuation V for $\mathcal{Q}$ over a universe U, some facts over U will always appear on this node responsible for V as well.

Definition 4.18. Let V be a valuation for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe U. A fact f over U is in $impFacts(V, \mathcal{Q})$ if and only if for every distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct and for every node $\kappa \in \mathcal{N}$, if $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $f \in rfacts_{\boldsymbol{P}}(\kappa)$.

Intuitively, we define $impFacts(V, \mathcal{Q})$ as the set of facts over U that appear on the node responsible for V under every distribution policy $\boldsymbol{P}$ over U under which $\mathcal{Q}$ is parallel-correct. Notice that by definition it trivially

follows that $V(body_{\mathcal{Q}}) \subseteq impFacts(V, \mathcal{Q})$ for every valuation V for a conjunctive query $\mathcal{Q}$.

Next, we describe some interesting properties of the set $impFacts(V, \mathcal{Q})$. These properties will prove useful to formulate a set of inference rules for $impFacts(V, \mathcal{Q})$.

Proposition 4.19. *Let V_1 and V_2 be valuations for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe U. If $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$, then $impFacts(V_1, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$.*

Proof. Assume valuations V_1 and V_2 as in the proposition. We prove that $impFacts(V_1, \mathcal{Q})$ equals $impFacts(V_2, \mathcal{Q})$ by showing that *(i)* for every fact $f \in impFacts(V_1, \mathcal{Q})$, it holds that $f \in impFacts(V_2, \mathcal{Q})$ and *(ii)* for every fact $f \in impFacts(V_2, \mathcal{Q})$, it holds that $f \in impFacts(V_1, \mathcal{Q})$.

(i) Let f be a fact in $impFacts(V_1, \mathcal{Q})$. For every distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct it holds by definition that:

1. for every node $\kappa \in \mathcal{N}$, if $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $f \in rfacts_{\boldsymbol{P}}(\kappa)$, since $f \in impFacts(V_1, \mathcal{Q})$, and
2. for every node $\kappa \in \mathcal{N}$, if $V_2(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, since $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$.

By combining these two statements, we get that for every node $\kappa \in \mathcal{N}$, if $V_2(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $f \in rfacts_{\boldsymbol{P}}(\kappa)$. We conclude that $f \in impFacts(V_2, \mathcal{Q})$.

(ii) Let f be a fact in $impFacts(V_2, \mathcal{Q})$. By definition, for every distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct it holds that:

1. for every node $\kappa \in \mathcal{N}$, if $V_2(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $f \in rfacts_{\boldsymbol{P}}(\kappa)$, since $f \in impFacts(V_2, \mathcal{Q})$, and
2. for every node $\kappa \in \mathcal{N}$, if $V_2(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, since $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$.

By Proposition 4.6, there is exactly one node κ_2 responsible for V_2. By applying the two previous statements, we get that $f \in rfacts_{\boldsymbol{P}}(\kappa_2)$ and $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$. Since $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$, there is furthermore no node $\kappa' \in \mathcal{N}$ different from κ_2 with $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$, as this would contradict our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. Since $f \in rfacts_{\boldsymbol{P}}(\kappa_2)$, we conclude that for every node $\kappa \in \mathcal{N}$, if $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, then $f \in rfacts_{\boldsymbol{P}}(\kappa)$. As a result, $f \in impFacts(V_1, \mathcal{Q})$. □

Proposition 4.20. *Let V_1 and V_2 be valuations for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe U. $impFacts(V_1, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$ if and only if there exists a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$.*

Proof. (if) Assume there is a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$. By Proposition 4.19, it follows that $impFacts(V, \mathcal{Q}) = impFacts(V_1, \mathcal{Q})$ and $impFacts(V, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$. We conclude that $impFacts(V_1, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$.

(only if) Assume $impFacts(V_1, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$. Notice that by definition $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$. It clearly follows that $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$. We conclude that there is a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$. □

We next present the following rules of inference deriving $impFacts(V, \mathcal{Q})$ given a valuation V for a conjunctive query with inequalities $\mathcal{Q}$ over a universe U:

Definition 4.21. Let V be a valuation for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe U. The rules of inference for $impFacts(V, \mathcal{Q})$ are:

1. If f is a fact in $V(body_{\mathcal{Q}})$, then $f \in impFacts(V, \mathcal{Q})$.
2. Let f be a fact over U. If there are two valuations V_1 and V_2 for $\mathcal{Q}$ over U such that $f \in impFacts(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \subseteq impFacts(V, \mathcal{Q})$, then $f \in impFacts(V, \mathcal{Q})$.

Notice that the second rule intuitively merges two sets $impFacts(V_1, \mathcal{Q})$ and $impFacts(V, \mathcal{Q})$ from as soon as they have the required facts for a valuation V_2 in common. This construction closely resembles Proposition 4.20, stating that $impFacts(V_1, \mathcal{Q})$ and $impFacts(V, \mathcal{Q})$ are the same if they have the required facts for a valuation V_2 in common.

Based on these rules of inference, we formulate an iterative approach to compute $impFacts(V, \mathcal{Q})$ for every valuation V for a given conjunctive query with inequalities $\mathcal{Q}$. Let $impF_i(V, \mathcal{Q})$ denote the computed set of facts in $impFacts(V, \mathcal{Q})$ after iteration i with $i \geq 0$. Let $\Delta impF_i(V, \mathcal{Q})$ denote the set of facts that were added to $impF_i(V, \mathcal{Q})$ in iteration i.

For each valuation V, the iteration scheme based on the rules of inference

now looks as follows:

$$
\begin{aligned}
impF_0(V, \mathcal{Q}) &= V(body_{\mathcal{Q}}) \\
impF_{i+1}(V, \mathcal{Q}) &= impF_i(V, \mathcal{Q}) \cup \left[\bigcup_{V_1 \in \mathcal{V}(V,i)} impF_i(V_1, \mathcal{Q}) \right] \\
\Delta impF_0(V, \mathcal{Q}) &= impF_0(V, \mathcal{Q}) \\
\Delta impF_{i+1}(V, \mathcal{Q}) &= impF_{i+1}(V, \mathcal{Q}) \setminus impF_i(V, \mathcal{Q})
\end{aligned}
$$

In this iteration scheme, $\mathcal{V}(V, i)$ is a shorthand notation for

$$\{V_1 \mid \exists V_2 : V_2(body_{\mathcal{Q}}) \subseteq impF_i(V_1, \mathcal{Q}) \wedge V_2(body_{\mathcal{Q}}) \subseteq impF_i(V, \mathcal{Q})\}.$$

Notice that the first rule of inference is applied to $impF_0(V, \mathcal{Q})$, while the second rule is applied in the computation of $impF_{i+1}(V, \mathcal{Q})$. We use this iteration scheme to prove that the given rules of inference are both sound and complete.

Proposition 4.22. *For a given valuation V for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe U, the rules of inference for $impFacts(V, \mathcal{Q})$ are sound.*

Proof. The proof is by induction on the iteration scheme. For each iteration $i \geq 0$, we show that $impF_i(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$.

(Base case) For $i = 0$, we get $impF_0(V, \mathcal{Q}) = V(body_{\mathcal{Q}})$. As $V(body_{\mathcal{Q}}) \subseteq impFacts(V, \mathcal{Q})$, it immediately follows that $impF_0(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$.

(Inductive step) Assume $impF_i(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$ for every valuation V for $\mathcal{Q}$ with $i \geq 0$. We prove that $impF_{i+1}(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$.

The set $impF_{i+1}(V, \mathcal{Q})$ consists of all the facts in $impF_i(V, \mathcal{Q})$, as well as the facts in $impF_i(V_1, \mathcal{Q})$ for each valuation V_1 for which there exists a valuation V_2 such that $V_2(body_{\mathcal{Q}}) \subseteq impF_i(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \subseteq impF_i(V, \mathcal{Q})$. We show that both *(i)* $impF_i(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$ and *(ii)* $impF_i(V_1, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$, thereby proving that $impF_{i+1}(V, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$. Notice that *(i)* is trivial by assumption, so we only focus on *(ii)* in the rest of this proof.

Let V_1 and V_2 be valuations for $\mathcal{Q}$ with $V_2(body_{\mathcal{Q}}) \subseteq impF_i(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \subseteq impF_i(V, \mathcal{Q})$. It follows by assumption that $V_2(body_{\mathcal{Q}}) \subseteq impFacts(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \subseteq impFacts(V, \mathcal{Q})$. According to Proposition 4.20, we know that $impFacts(V_1, \mathcal{Q})$ equals $impFacts(V, \mathcal{Q})$. Because $impF_i(V_1, \mathcal{Q}) \subseteq impFacts(V_1, \mathcal{Q})$ and $impFacts(V_1, \mathcal{Q}) = impFacts(V, \mathcal{Q})$, we conclude that $impF_i(V_1, \mathcal{Q}) \subseteq impFacts(V, \mathcal{Q})$. □

Assume a finite set of data values $\mathbf{dom}_k \subseteq \mathbf{dom}$. Notice that for every conjunctive query with inequalities $\mathcal{Q}$ over a database scheme $\mathcal{D}$ and for every universe $U \subseteq \mathbf{dom}_k$, both $facts(\mathcal{D}, U)$ and the number of valuations

for $\mathcal{Q}$ over U are finite as well. Therefore, it can easily be seen that for every conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and universe $U \subseteq \mathbf{dom}_k$, there has to be an iteration i such that $\Delta impF_i(V, \mathcal{Q})$ is empty for every valuation V for $\mathcal{Q}$ over U. Notice furthermore that for every iteration $j \geq i$ the set $\Delta impF_j(V, \mathcal{Q})$ is empty as well, as the iteration scheme only uses facts from the previous iteration to generate facts for the next iteration. Thus, if there was no change during the previous iteration, there will certainly be no change in the next iteration.

We conclude that for every conjunctive query $\mathcal{Q}$ over a database scheme $\mathcal{D}$ and for every universe $U \subseteq \mathbf{dom}_k$, the iteration scheme eventually reaches a point where the computed set of facts $impF_i(V, \mathcal{Q})$ will no longer change for every valuation V for $\mathcal{Q}$ over U. We refer to this set as $impF(V, \mathcal{Q})$. Based on this set $impF(V, \mathcal{Q})$, we prove that the rules of inference are complete. To this end, we first state the following lemma, closely related to Proposition 4.20.

Lemma 4.23. *Let V_1 and V_2 be valuations for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe $U \subseteq \mathbf{dom}_k$. $impF(V_1, \mathcal{Q}) = impF(V_2, \mathcal{Q})$ if and only if there exists a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impF(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impF(V_2, \mathcal{Q})$.*

Proof. (if) Let V be a valuation for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impF(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impF(V_2, \mathcal{Q})$. Assume $impF(V_1, \mathcal{Q}) \neq impF(V_2, \mathcal{Q})$. Let i denote the iteration after which the iteration scheme remains the same. Thus, $impF_i(V_1, \mathcal{Q}) = impF_{i+1}(V_1, \mathcal{Q}) = impF(V_1, \mathcal{Q})$ and analogously $impF_i(V_2, \mathcal{Q}) = impF_{i+1}(V_2, \mathcal{Q}) = impF(V_2, \mathcal{Q})$.

Since $impF_i(V_1, \mathcal{Q}) \neq impF_i(V_2, \mathcal{Q})$, there exists a fact f with:

- $f \in impF_i(V_1, \mathcal{Q})$ and $f \notin impF_i(V_2, \mathcal{Q})$, or
- $f \notin impF_i(V_1, \mathcal{Q})$ and $f \in impF_i(V_2, \mathcal{Q})$.

Assume the first case is true (if instead the second case would be true, we just need to swap V_1 and V_2). Since $V(body_{\mathcal{Q}})$ is a subset of or equal to both $impF_i(V_1, \mathcal{Q})$ and $impF_i(V_2, \mathcal{Q})$, it immediately follows by definition of $impF_{i+1}(V_2, \mathcal{Q})$ that all the facts in $impF_i(V_1, \mathcal{Q})$ are in $impF_{i+1}(V_2, \mathcal{Q})$. This implies that $f \in impF_{i+1}(V_2, \mathcal{Q})$, thereby contradicting our earlier statement that $impF_i(V_2, \mathcal{Q})$ equals $impF_{i+1}(V_2, \mathcal{Q})$. We conclude that our initial assumption cannot hold, thus $impF(V_1, \mathcal{Q}) = impF(V_2, \mathcal{Q})$

(only if) Assume $impF(V_1, \mathcal{Q})$ and $impF(V_2, \mathcal{Q})$ are equal. By definition of $impF_0(V_1, \mathcal{Q})$, we know for sure that $V_1(body_{\mathcal{Q}}) \subseteq impF(V_1, \mathcal{Q})$. As a result, there surely exists a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impF(V_1, \mathcal{Q})$ and $V(body_{\mathcal{Q}}) \subseteq impF(V_2, \mathcal{Q})$, namely V_1. □

By negating both parts of Lemma 4.23, we get the following corollary:

Corollary 4.24. *Let V_1 and V_2 be valuations for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe $U \subseteq \boldsymbol{dom}_k$. $impF(V_1, \mathcal{Q}) \neq impF(V_2, \mathcal{Q})$ if and only if there is no valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq impF(V_1, \mathcal{Q}) \cap impF(V_2, \mathcal{Q})$.*

The previous corollary will prove useful to show that the rules of inference for $impFacts(V, \mathcal{Q})$ are complete.

Proposition 4.25. *For a given valuation V for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ over a universe $U \subseteq \boldsymbol{dom}_k$, the rules of inference for $impFacts(V, \mathcal{Q})$ are complete.*

Proof. Let V be a valuation for a query $\mathcal{Q}$ over a universe $U \subseteq \mathbf{dom}_k$. Based on the iteration scheme, we show that the rules of inference are complete by proving by contraposition that for every fact $f \in impFacts(V, \mathcal{Q})$ it also holds that $f \in impF(V, \mathcal{Q})$.

To this end, assume a fact $f \notin impF(V, \mathcal{Q})$. We show that f does not appear in $impFacts(V, \mathcal{Q})$ by constructing a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct and for which $f \notin rfacts_{\boldsymbol{P}}(\kappa)$ if $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ for every node $\kappa \in \mathcal{N}$.

The construction of $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ uses a total function map to map each valuation V' for $\mathcal{Q}$ over U onto a node $\kappa \in \mathcal{N}$. We require that for every pair of valuations V_i and V_j for $\mathcal{Q}$ over U, $map(V_i) = map(V_j)$ if and only if $impF(V_i, \mathcal{Q}) = impF(V_j, \mathcal{Q})$. Since the number of valuations is finite, $\mathcal{N}$ is finite as well.[6]

For every valuation V_i for $\mathcal{Q}$ over U, we define $rfacts_{\boldsymbol{P}}(map(V_i)) = impF(V_i, \mathcal{Q})$. Notice that $rfacts_{\boldsymbol{P}}$ is well-defined, even if multiple valuations V_i and V_j are assigned to the same node κ, as in this case it is required that $impF(V_i, \mathcal{Q}) = impF(V_j, \mathcal{Q})$.

We now prove that $\mathcal{Q}$ is parallel-correct under the constructed policy $\boldsymbol{P}$ by showing that for every valuation V' for $\mathcal{Q}$ over U, there is exactly one node κ with $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Observe that there trivially is at least one node κ with $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, namely $map(V')$. Therefore, it suffices to show that there is no node $\kappa \neq map(V')$ with $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

Assume there is a node $\kappa \neq map(V')$ with $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. By construction, there is a valuation V'' with $map(V'') = \kappa$ and $V'(body_{\mathcal{Q}}) \subseteq impF(V'', \mathcal{Q})$. Since $map(V') \neq map(V'')$, we know that $impF(V', \mathcal{Q}) \neq impF(V'', \mathcal{Q})$. According to Corollary 4.24, there is no valuation V with $V(body_{\mathcal{Q}}) \subseteq impF(V', \mathcal{Q}) \cap impF(V'', \mathcal{Q})$. But this is a contradiction, since $V'(body_{\mathcal{Q}}) \subseteq impF(V', \mathcal{Q}) \cap impF(V'', \mathcal{Q})$. We conclude that there cannot be a node $\kappa \neq map(V')$ with $V'(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

To conclude our proof, we need to show that for every node $\kappa \in \mathcal{N}$ it holds that $f \notin rfacts_{\boldsymbol{P}}(\kappa)$ if $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Notice that $\kappa_V = map(V)$ is the only node with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_V)$, so it suffices to show that $f \notin rfacts_{\boldsymbol{P}}(\kappa_V)$. But this is trivial since $rfacts_{\boldsymbol{P}}(\kappa_V) = impF(V, \mathcal{Q})$ and $f \notin impF(V, \mathcal{Q})$. □

[6] We implicitly assume that $\mathcal{N}$ contains no nodes that aren't reachable by map.

4.3.3 A characterization for parallel-correctness transfer

The notion of $\mathit{impFacts}(V, \mathcal{Q})$ is used in the following condition, which is both necessary and sufficient for transferability under bag semantics:

Condition 4.26. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. For each valuation V' for $\mathcal{Q}'$ over a universe U there exists a valuation V for $\mathcal{Q}$ over U such that $V(\mathit{body}_{\mathcal{Q}}) \subseteq V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V, \mathcal{Q})$.*

Before proving that Condition 4.26 is indeed necessary and sufficient for transferability, we first consider a consequence when Condition 4.26 is not satisfied, witnessed by some valuation V' for $\mathcal{Q}'$. More specifically, we claim that in this case there is no valuation V for $\mathcal{Q}$ with $V(\mathit{body}_{\mathcal{Q}}) \subseteq V'(\mathit{body}_{\mathcal{Q}'})$ or there is no valuation V for $\mathcal{Q}$ with $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V, \mathcal{Q})$. Although this claim might seem like a trivial consequence directly obtainable by applying set logic, it is not.

Example 4.27. To show that our claim is not a direct result of set logic, we construct a counterexample based on sets instead of valuations and an arbitrary function f instead of *impFacts*.

Consider two sets of sets $\mathcal{A}$ and $\mathcal{B}$ with

$$\mathcal{A} = \{\{a\}, \{b, c\}\},$$

and

$$\mathcal{B} = \{\{a, b\}\}.$$

Furthermore, let f be a function having $\mathcal{A}$ as the domain with

$$f(\{a\}) = \{a, c\},$$
$$f(\{b, c\}) = \{a, b, c\}.$$

An analogous reformulation of Condition 4.26 based on these sets and the function f is as follows: for every set B in $\mathcal{B}$, there is a set A in $\mathcal{A}$ with $A \subseteq B \subseteq f(A)$. This condition clearly does not hold, since

$$\{a\} \quad \subseteq \{a, b\} \not\subseteq \{a, c\}, \text{ and}$$
$$\{b, c\} \not\subseteq \{a, b\} \subseteq \{a, b, c\}.$$

Let $B = \{a, b\}$ be the set in $\mathcal{B}$ witnessing the failure of the condition. A reformulation of our claim is now as follows: there is no set A in $\mathcal{A}$ with $A \subseteq B$ or there is no set A in $\mathcal{A}$ with $B \subseteq f(A)$. It can easily be seen that this claim does not hold in this example. We conclude that our claim does not hold in general for arbitrary sets. ■

Lemma 4.28. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. Condition 4.26 isn't satisfied, witnessed by some valuation V' for $\mathcal{Q}'$ over a universe U, if and only if there is no valuation V for $\mathcal{Q}$ over U with $V(\mathit{body}_{\mathcal{Q}}) \subseteq V'(\mathit{body}_{\mathcal{Q}'})$ or there is no valuation V for $\mathcal{Q}$ over U with $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V, \mathcal{Q})$.*

Proof. (if) Let V' be a valuation as described in Lemma 4.28. For both cases, it immediately follows that there cannot exist a valuation V for $\mathcal{Q}$ over U With $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. Therefore, Condition 4.26 isn't satisfied.

(only if) Assume Condition 4.26 isn't satisfied, witnessed by some valuation V' for $\mathcal{Q}'$ over U. By assumption, every valuation V for $\mathcal{Q}$ over U satisfies one of the following three conditions:

1. $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \not\subseteq impFacts(V, \mathcal{Q})$
2. $V(body_{\mathcal{Q}}) \not\subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$
3. $V(body_{\mathcal{Q}}) \not\subseteq V'(body_{\mathcal{Q}'}) \not\subseteq impFacts(V, \mathcal{Q})$

Notice that condition 1 and condition 2 cannot occur together. Indeed, assume there are two valuations V_1 and V_2 for $\mathcal{Q}$ over U satisfying respectively condition 1 and condition 2. That is, $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \not\subseteq impFacts(V_1, \mathcal{Q})$ and $V_2(body_{\mathcal{Q}}) \not\subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V_2, \mathcal{Q})$. It immediately follows that $V_1(body_{\mathcal{Q}}) \subseteq impFacts(V_2, \mathcal{Q})$. According to Proposition 4.19, $impFacts(V_1, \mathcal{Q}) = impFacts(V_2, \mathcal{Q})$. But this creates a contradiction, as $V'(body_{\mathcal{Q}'}) \not\subseteq impFacts(V_1, \mathcal{Q})$ and $V'(body_{\mathcal{Q}'}) \subseteq impFacts(V_2, \mathcal{Q})$.

If every valuation V for $\mathcal{Q}$ over U satisfies condition 2 or condition 3, there is no valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. Analogously, if every valuation V for $\mathcal{Q}$ over U satisfies condition 1 or condition 3, there is no valuation V for $\mathcal{Q}$ over U with $V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. □

We use Lemma 4.28 to prove that Condition 4.26 is necessary and sufficient for transferability under bag semantics.

Proposition 4.29. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. Parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if Condition 4.26 is satisfied.*

Proof. (if) Assume Condition 4.26 is satisfied. Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy under which $\mathcal{Q}$ is parallel-correct. We prove that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well. According to Proposition 4.6, it suffices to show that for each valuation V' for $\mathcal{Q}'$ over U there is exactly one node κ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. By assumption, there is a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. Since $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$, there is exactly one node κ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

We show that κ is the only node responsible for V'. To this end, we separately prove that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$ and that there is no node κ' different from κ with $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$.

Since $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ and since $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$, the set of facts $impFacts(V, \mathcal{Q})$ contains by definition only facts that are guaranteed to be present on κ as well, or more formally $impFacts(V, \mathcal{Q}) \subseteq$

$rfacts_{\boldsymbol{P}}(\kappa)$. It now trivially follows from $V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$ that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

Since $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$, there cannot be a node κ' different from κ with $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$, as this would trivially imply $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$, thereby contradicting our assumption that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$.

(only if) The proof is by contraposition. Assume Condition 4.26 isn't satisfied, witnessed by some valuation V' for $\mathcal{Q}'$ over a universe U. We prove that parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$.

According to Lemma 4.28, there is no valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$ or there is no valuation V for $\mathcal{Q}$ over U with $V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. We next consider both cases separately. In both cases, we construct a distribution policy $\boldsymbol{P}$ under which $\mathcal{Q}$ is parallel-correct, but $\mathcal{Q}'$ is not. This policy $\boldsymbol{P}$ implies that parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$.

First, consider the case where there is no valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$. In this case, $\boldsymbol{P}$ is constructed as follows over a network with two nodes κ_1 and κ_2: all facts are assigned to κ_1 and the facts in $V'(body_{\mathcal{Q}'})$ are assigned to κ_2 as well. Notice that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$, as the required facts for each valuation V for $\mathcal{Q}$ only meet at κ_1. Indeed, if the required facts for a valuation V for $\mathcal{Q}$ would meet at κ_2 then this would imply that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'})$, contradicting our assumption that no such valuation V exists. The required facts for V' however trivially meet at both nodes. By Proposition 4.6, $\mathcal{Q}'$ is not parallel-correct under $\boldsymbol{P}$. We conclude that in this case parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$.

Next, consider the case where there is no valuation V for $\mathcal{Q}$ over U with $V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. By definition of $impFacts(V, \mathcal{Q})$, for each valuation V for $\mathcal{Q}$ over U, there is a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ under which $\mathcal{Q}$ is parallel-correct such that $V'(body_{\mathcal{Q}'}) \not\subseteq rfacts_{\boldsymbol{P}}(\kappa)$, with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. In other words, for each valuation V for $\mathcal{Q}$, there is a distribution policy $\boldsymbol{P}$ that does not map all the required facts for V' on the same node as the required facts for V.

Furthermore, we can safely assume the existence of a valuation V_1 for $\mathcal{Q}$ over U with $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}})$. Indeed, if this valuation V_1 would not exist, these queries would be covered by the first case. By assumption, $V'(body_{\mathcal{Q}'}) \not\subseteq impFacts(V_1, \mathcal{Q})$. Thus, there is a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ under which $\mathcal{Q}$ is parallel-correct and $V'(body_{\mathcal{Q}'}) \not\subseteq rfacts_{\boldsymbol{P}}(\kappa)$, with $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. It now suffices to show that $\mathcal{Q}'$ is not parallel-correct under this distribution policy $\boldsymbol{P}$.

Assume $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$. since $V'(body_{\mathcal{Q}'}) \not\subseteq rfacts_{\boldsymbol{P}}(\kappa)$, there has to be another node κ' different from κ such that $V'(body_{\mathcal{Q}'}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$. However, since $V_1(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}})$, it immediately follows

that $V_1(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$. But this would imply that V_1 is satisfiable on two different nodes κ and κ'. According to Proposition 4.6, $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. This contradiction implies that $\mathcal{Q}'$ is not parallel-correct under $\boldsymbol{P}$. We conclude that in this case, parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$. □

4.4 Parallel-correctness transfer complexity

In this section, we focus on the complexity of parallel-correctness transfer. More specifically, we study the complexity of the problem **PC-Trans**$(\mathcal{C}, \mathcal{C}')$, with $\mathcal{C}$ and $\mathcal{C}'$ a query class.

	PC-Trans$(\mathcal{C}, \mathcal{C}')$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	query $\mathcal{Q}' \in \mathcal{C}'$
Question:	Does parallel-correctness transfer from $\mathcal{Q}$ to $\mathcal{Q}'$?

4.4.1 Equivalence of the characterization over infinite and finite domains

To determine the complexity of parallel-correctness for various query classes, we use the characteristic described in Proposition 4.29. Direct use of this characteristic is not feasible, as it would require checking an infinite number of valuations over an infinite domain **dom**. It is however possible to limit our domain to a finite domain $\mathbf{dom}_k$, as long as the number of values in this domain is at least as much as the number of values a valuation can use for one of both input queries. This property is stated in more detail in the following claim:

Claim 4.30. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{dom}_k = \{1, \ldots, k\}$ be a finite subset of **dom**, where $k = max(varmax(\mathcal{Q}), varmax(\mathcal{Q}'))$. The following conditions are equivalent:*

(1) For each valuation V' for $\mathcal{Q}'$ over a universe $U \subseteq \boldsymbol{dom}$, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$.

(2) For each valuation W' for $\mathcal{Q}'$ over a universe $U_k \subseteq \boldsymbol{dom}_k$, there exists a valuation W for $\mathcal{Q}$ over U_k such that $W(body_{\mathcal{Q}}) \subseteq W'(body_{\mathcal{Q}'}) \subseteq impFacts(W, \mathcal{Q})$.

Proof. Condition (1) implies Condition (2): This implication is trivial, since $\mathbf{dom}_k$ is a finite subset of **dom**. This implies that every universe $U_k \subseteq \mathbf{dom}_k$ is a universe over **dom** as well.

Condition (2) implies Condition (1): The proof is by contraposition. Assume Condition (1) does not hold. We prove that Condition (2) doesn't hold. Since Condition (1) doesn't hold, there is a valuation V'_U for $\mathcal{Q}'$ over a universe U such that there is no valuation V_U for $\mathcal{Q}$ over U with $V_U(\mathit{body}_\mathcal{Q}) \subseteq V'_U(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V_U, \mathcal{Q})$. According to Lemma 4.28, there are two possible cases: *(1)* For every valuation V_U for $\mathcal{Q}$ over U, it holds that $V_U(\mathit{body}_\mathcal{Q}) \not\subseteq V'_U(\mathit{body}_{\mathcal{Q}'})$, or *(2)* for every valuation V_U for $\mathcal{Q}$ over U, it holds that $V'_U(\mathit{body}_{\mathcal{Q}'}) \not\subseteq \mathit{impFacts}(V_U, \mathcal{Q})$.

Now consider the universe $U' \stackrel{\text{def}}{=} \mathit{adom}(V'_U(\mathit{body}_\mathcal{Q}))$. Observe that, by definition of U', the valuation V'_U is a valid valuation for $\mathcal{Q}'$ over U' as well, denoted by $V'_{U'}$. Next, we prove that both cases remain true when considering U' instead of U. The first case is trivial, as $U' \subseteq U$. If there is no valuation V_U for $\mathcal{Q}$ over U with $V_U(\mathit{body}_\mathcal{Q}) \subseteq V'_U(\mathit{body}_{\mathcal{Q}'})$, then there will surely be no valuation $V_{U'}$ for $\mathcal{Q}$ over U' such that $V_{U'}(\mathit{body}_\mathcal{Q}) \subseteq V'_{U'}(\mathit{body}_{\mathcal{Q}'})$.

In order to prove the second case over U', consider a valuation $V_{U'}$ for $\mathcal{Q}$ over U'. We need to prove that $V'_{U'}(\mathit{body}_{\mathcal{Q}'}) \not\subseteq \mathit{impFacts}(V_{U'}, \mathcal{Q})$. Since U' is a subset of U, this valuation $V_{U'}$ is a valid valuation for $\mathcal{Q}$ over U as well, denoted by V_U. By assumption, $V'_U(\mathit{body}_{\mathcal{Q}'}) \not\subseteq \mathit{impFacts}(V_U, \mathcal{Q})$ when valuated over U. This implies the existence of a distribution policy $\boldsymbol{P}$ over U under which $\mathcal{Q}$ is parallel-correct. Furthermore, $\boldsymbol{P}$ does not map all the required facts for V'_U on the same node as the required facts for V_U. Based on this distribution policy $\boldsymbol{P}$ over U, we can now easily construct a distribution policy $\boldsymbol{P'}$ over U' in such a way that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P'}$ as well. To construct $\boldsymbol{P'}$ from $\boldsymbol{P}$, we only need to remove the facts using data values that are not in U'. All facts over U' thus are still mapped onto the same nodes as in $\boldsymbol{P}$. It can easily be seen that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P'}$, as every valuation for $\mathcal{Q}$ over U' clearly is still mapped onto exactly one node. Notice that, just like $\boldsymbol{P}$, this distribution policy $\boldsymbol{P'}$ does not map all the required facts for $V'_{U'}$ on the same node as the required facts for $V_{U'}$. Consequently, $V'_{U'}(\mathit{body}_{\mathcal{Q}'}) \not\subseteq \mathit{impFacts}(V_{U'}, \mathcal{Q})$ when using U' as a universe.

Since both cases remain true over U', we conclude that there is no valuation $V_{U'}$ for $\mathcal{Q}$ over U' with $V_{U'}(\mathit{body}_\mathcal{Q}) \subseteq V'_{U'}(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V_{U'}, \mathcal{Q})$.

To conclude our proof, we now show that Condition (2) doesn't hold. Let $\pi : \mathbf{dom} \to \mathbf{dom}$ be an arbitrary bijection with $U_k \stackrel{\text{def}}{=} \pi(U') \subseteq \mathbf{dom}_k$. Furthermore, let $W' \stackrel{\text{def}}{=} \pi \circ V'_{U'}$ be a valuation for $\mathcal{Q}'$ over U_k. There cannot be a valuation W for $\mathcal{Q}$ over U_k such that $W(\mathit{body}_\mathcal{Q}) \subseteq W'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(W, \mathcal{Q})$. Indeed, if this valuation W for $\mathcal{Q}$ over U_k would exist, the valuation $V_{U'} \stackrel{\text{def}}{=} \pi^{-1} \circ W$ would be a valuation for $\mathcal{Q}$ over U' with $V_{U'}(\mathit{body}_\mathcal{Q}) \subseteq V'_{U'}(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V_{U'}, \mathcal{Q})$, thereby contradicting our earlier conclusion that no such $V_{U'}$ exists. □

4.4.2 Conjunctive queries

Proposition 4.31. *The problem* **PC-Trans**$(\mathbf{CQ}^{\neq}, \mathbf{CQ}^{\neq})$ *is in* EXPTIME.

Proof. The proof is by construction of an EXPTIME algorithm deciding **PC-Trans**$(\mathbf{CQ}^{\neq}, \mathbf{CQ}^{\neq})$ on input $\mathcal{Q}, \mathcal{Q}' \in \mathbf{CQ}^{\neq}$. Recall from Claim 4.30 that we only need to check transferability over a finite domain $\mathbf{dom}_k$ containing k different values, where $k = max(varmax(\mathcal{Q}), varmax(\mathcal{Q}'))$.

In general, the algorithm iterates over every universe $U \subseteq \mathbf{dom}_k$. For every such universe U, the following two steps are performed:

1. The set $\mathit{impFacts}(V, \mathcal{Q})$ is calculated for every valuation V for $\mathcal{Q}$ over the universe U.

2. For every valuation V' for $\mathcal{Q}'$ over U it is checked whether there is a valuation V for $\mathcal{Q}$ over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq \mathit{impFacts}(V, \mathcal{Q})$.

If the check in the second step succeeds for every universe U and for every valuation V' for $\mathcal{Q}'$ over U, the algorithm returns *true*, indicating that parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. Otherwise, *false* is returned.

Let n denote the size of the input $\langle \mathcal{Q}, \mathcal{Q}' \rangle$. Observe that the size k of $\mathbf{dom}_k$ is by definition linear in function of n. As a result, the number of possible universes $U \subseteq \mathbf{dom}_k$ is exponential in function of the input size n. Therefore, it is possible to enumerate all universes $U \subseteq \mathbf{dom}_k$ in exponential time. We next focus on such a universe $U \subseteq \mathbf{dom}_k$ and show that both steps described above can be executed in exponential time, thereby proving that the algorithm is indeed in EXPTIME.

The computation of $\mathit{impFacts}(V, \mathcal{Q})$ for every valuation V for $\mathcal{Q}$ over U is based on the rules of inference defined in Definition 4.21. First, according to the first rule, $\mathit{impFacts}(V, \mathcal{Q})$ is initialized with $V(body_{\mathcal{Q}})$. The second rule is applied as follows: for every triple of valuations V_1, V_2, V_3 for $\mathcal{Q}$ over U, merge $\mathit{impFacts}(V_1, \mathcal{Q})$ and $\mathit{impFacts}(V_2, \mathcal{Q})$ if $V_3(body_{\mathcal{Q}}) \subseteq \mathit{impFacts}(V_1, \mathcal{Q})$ and $V_3(body_{\mathcal{Q}}) \subseteq \mathit{impFacts}(V_2, \mathcal{Q})$. This second step is repeated until no new merges are possible.

Since the number of variables appearing in $\mathcal{Q}$ is linear in function of n and since valuations for $\mathcal{Q}$ over U are defined as mappings from variables appearing in $\mathcal{Q}$ onto values in U, we conclude that the total number of different valuations for $\mathcal{Q}$ over U is exponential in function of n. Therefore, iterating over all triples of valuations V_1, V_2, V_3 for $\mathcal{Q}$ over U can be done in exponential time. The number of possible merges is furthermore exponential in function of n as well, leading to the conclusion that the computation of $\mathit{impFacts}(V, \mathcal{Q})$ for every valuation V for $\mathcal{Q}$ over U can be performed in exponential time.

The second step is trivial: iterate over all the valuations V' for $\mathcal{Q}'$ over U and check for each such valuation V' if there exists a valuation V for $\mathcal{Q}$

over U with $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. Analogously to the number of valuations for $\mathcal{Q}$ over U, the number of possible valuations for $\mathcal{Q}'$ over U is exponential in function of n. Therefore, this second step uses exponential time as well. □

4.4.3 Conjunctive queries without self-joins

The difficult part in applying the characterization for transferability is the computation of the set $impFacts(V, \mathcal{Q})$ for a given valuation V for a query $\mathcal{Q}$ over a universe U. Determining whether a fact f is in $impFacts(V, \mathcal{Q})$ can be simplified if $\mathcal{Q}$ is further constrained. In this section, we focus on conjunctive queries with inequalities without self-joins, denoted $\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$. Consider a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$ with $body_{\mathcal{Q}} = \{R_1(y_1), \ldots, R_m(y_m)\}$. Since $\mathcal{Q}$ has no self-joins, every atom in $body_{\mathcal{Q}}$ has a different relation name. In other words, $R_i \neq R_j$ for $1 \leq i < j \leq m$.

Proposition 4.32. *Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$. For every valuation V for $\mathcal{Q}$ over a universe U, $impFacts(V, \mathcal{Q}) = V(body_{\mathcal{Q}})$.*

Proof. Consider a valuation V over a universe U for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$ with $body_{\mathcal{Q}} = \{R_1(y_1), \ldots, R_m(y_m)\}$. We prove that a fact f over U is in $impFacts(V, \mathcal{Q})$ if and only if $f \in V(body_{\mathcal{Q}})$.

(if) Assume $f \in V(body_{\mathcal{Q}})$. For every distribution policy $\boldsymbol{P}$ over U under which $\mathcal{Q}$ is parallel-correct, f trivially appears on the same node as the facts in $V(body_{\mathcal{Q}})$. Consequently, $f \in impFacts(V, \mathcal{Q})$.

(only if) The proof is by contraposition. Assume $f \notin V(body_{\mathcal{Q}})$. We show that $f \notin impFacts(V, \mathcal{Q})$ by constructing a distribution policy $\boldsymbol{P}$ over U such that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ and f does not appear on the node responsible for V.

Let $S = \{S_1, \ldots, S_{2^m}\}$ be the powerset of $\{R_1, \ldots, R_m\}$. We now construct $\boldsymbol{P}$ over a network $\mathcal{N} = \{\kappa_1, \ldots, \kappa_{2^m}\}$ as follows: for every fact f over a relation R_j with $1 \leq j \leq m$ and for every node κ_i with $1 \leq i \leq 2^m$, we map f on κ_i if and only if one of the following two conditions is satisfied:

- $f \in V(body_{\mathcal{Q}})$ and $R_j \in S_i$, or
- $f \notin V(body_{\mathcal{Q}})$ and $R_j \notin S_i$.

Observe that by construction every combination of facts $f_1, \ldots, f_m$ with f_j a fact over R_j is present on exactly one node. Consequently, there is exactly one node κ for every valuation V' for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}}) \in rfacts_{\boldsymbol{P}}(\kappa)$. As a result, $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$.

Notice furthermore that the required facts for V are mapped onto a node κ_i with $S_i = \{R_1, \ldots, R_m\}$. By construction, this node contains only facts

in $V(body_{\mathcal{Q}})$. We conclude that a fact $f \notin V(body_{\mathcal{Q}})$ does not appear on the same node as all the facts in $V(body_{\mathcal{Q}})$ and, as a result, $f \notin impFacts(V, \mathcal{Q})$. □

Notice that the property behind conjunctive queries without self-joins causing $impFacts(V, \mathcal{Q})$ to always equal $V(body_{\mathcal{Q}})$ is the fact that the required facts for a valuation are never a strict subset of the required facts for another valuation. Therefore, it might be tempting to think that Proposition 4.32 could be further extended to strongly minimal queries, as their definition is based on the notion of containment of required facts for one valuation in another valuation as well. Unfortunately, this is not true, since strongly minimal queries only require that for every valuation V deriving an arbitrary fact f there is no other valuation V' deriving the same fact f while requiring strictly less facts. In other words, the required facts for a valuation V might still be a strict subset of the required facts for a valuation V', as long as V and V' do not derive the same fact.

Example 4.33. As a counterexample showing that for strongly minimal conjunctive queries $impFacts(V, \mathcal{Q})$ not necessarily equals $V(body_{\mathcal{Q}})$, consider the following conjunctive query $\mathcal{Q}$,

$$H(x, y, z) \leftarrow R(x, y), R(y, z).$$

Since every variable appears in the head atom, $\mathcal{Q}$ is a full conjunctive query and thus a strongly minimal conjunctive query as well. Notice that this query $\mathcal{Q}$ closely resembles the conjunctive query $\mathcal{Q}$ of Example 4.15, except the fact that variable y now appears in the head atom as well. The modification doesn't alter the reasoning in Example 4.15 showing that a distribution policy under which $\mathcal{Q}$ is parallel-correct has to map every fact onto the same node. We therefore conclude that $impFacts(V, \mathcal{Q})$ not necessarily equals $V(body_{\mathcal{Q}})$. ■

Proposition 4.32 allows us to reformulate Proposition 4.29 as follows:

Proposition 4.34. *Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$ and let $\mathcal{Q}'$ be a query in $\mathbf{CQ}^{\neq}$. Parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if for each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) = V'(body_{\mathcal{Q}'})$.*

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be as described in Proposition 4.34. According to Proposition 4.29, parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if for each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq impFacts(V, \mathcal{Q})$. Since $\mathcal{Q}$ is a query without self-joins, by Proposition 4.32, $impFacts(V, \mathcal{Q}) = V(body_{\mathcal{Q}})$. Thus, parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if for each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such

that $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$. But $V(body_{\mathcal{Q}}) \subseteq V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$ if and only if $V(body_{\mathcal{Q}}) = V'(body_{\mathcal{Q}'})$. Consequently, parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if for each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) = V'(body_{\mathcal{Q}'})$. □

We are now ready to give an improved upper bound on the time complexity of **PC-Trans**($\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}, \mathbf{CQ}^{\neq}$).

Proposition 4.35. *The problem* **PC-Trans**($\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}, \mathbf{CQ}^{\neq}$) *is in* Π^p_2.

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be the input queries for **PC-Trans**($\mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}, \mathbf{CQ}^{\neq}$). According to Proposition 4.34, it suffices to show that there is a Π^p_2-algorithm that checks if for each valuation V' for $\mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) = V'(body_{\mathcal{Q}'})$. Since $V(body_{\mathcal{Q}}) = V'(body_{\mathcal{Q}'})$ can obviously be checked in polynomial time, the construction of the Π^p_2-algorithm is trivial. □

Chapter 5

Relation between set and bag semantics

In this chapter, we focus on the relation between set and bag semantics when considering parallel-correctness and transferability. To facilitate the distinction between set and bag semantics, we denote that the query $\mathcal{Q}$ is parallel-correct under distribution policy $\boldsymbol{P}$ under set or bag semantics by respectively $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ and $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$. Analogously, we use $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ to indicate that parallel-correctness transfers from a query $\mathcal{Q}$ to a query $\mathcal{Q}'$ under respectively set and bag semantics.

5.1 Parallel-correctness

The characterizations for parallel-correctness for conjunctive queries with inequalities under set and bag semantics are quite different: while the former only focuses on minimal valuations over a conjunctive query $\mathcal{Q}$, the latter considers all valuations over $\mathcal{Q}$. Furthermore, the former requires that the required facts for a considered valuation V are mapped onto at least one node in the network, while the latter requires that these facts are mapped onto exactly one node. Intuitively, these restrictions for queries in $\mathbf{CQ}^{\neq}$ under set semantics are less strict than those under bag semantics. As a result, parallel-correctness for queries in $\mathbf{CQ}^{\neq}$ under bag semantics implies parallel-correctness for conjunctive queries under set semantics. This observation is formalised in the following proposition:

Proposition 5.1. *Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy. If $PC_{bag}(\boldsymbol{P}, \mathcal{Q})$, then $PC_{set}(\boldsymbol{P}, \mathcal{Q})$.*

Proof. The proof is by contraposition. Assume $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ does not hold for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$. We prove that $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ doesn't hold. By Proposition 3.8, there exists a

minimal valuation V for $\mathcal{Q}$ over U for which there is no node $\kappa \in \mathcal{N}$ with $V(\mathit{body}_\mathcal{Q}) \not\subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$.

According to Proposition 4.6, This valuation V is an immediate counterexample for $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$. Consequently, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ does not hold. □

The converse of Proposition 5.1 does not hold true, as the next example shows:

Example 5.2. For a counterexample showing that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ doesn't necessarily imply $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$, consider the conjunctive query $\mathcal{Q}$,

$$T(x) \leftarrow R(x), R(y),$$

and the network $\mathcal{N} = \{\kappa_1, \kappa_2\}$. Assume a universe $U = \{a, b\}$. Let $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ be a distribution policy over $\mathcal{N}$ with $\mathit{rfacts}_{\boldsymbol{P}}(\kappa_1) = \{R(a)\}$ and $\mathit{rfacts}_{\boldsymbol{P}}(\kappa_2) = \{R(b)\}$.

When focusing on the valuation $V = \{x \mapsto a, y \mapsto b\}$, we see that there is no node $\kappa \in \mathcal{N}$ such that $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. As a result, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ doesn't hold. $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ on the other hand does hold. Indeed, the required facts for each minimal valuation are present on a node. For example, V is not a minimal valuation, but the valuation $V' = \{x \mapsto a, y \mapsto a\}$ is a minimal valuation deriving the same fact as V, and $V'(\mathit{body}_\mathcal{Q}) = \{R(a)\} \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_1)$. ■

Example 5.2 illustrates one of the key differences between set and bag semantics. Under set semantics, nonminimal valuations don't need to be satisfied, as the related minimal valuation will derive the same fact with strictly less required facts. However, these nonminimal valuations play an important role under bag semantics because they influence the final multiplicity of the derived fact.

One might think that constraining the considered conjunctive queries to strongly minimal queries would be sufficient to let $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ imply $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$. Although this constraint would solve the issue described in Example 5.2, it is not sufficient to make the implication true. Furthermore, the following example shows that it might be very hard to come up with a restriction on the form of conjunctive queries alone to let parallel-correctness coincide for set and bag semantics.

Example 5.3. Consider the conjunctive query $\mathcal{Q}$,

$$T(x) \leftarrow R(x),$$

and the network $\mathcal{N} = \{\kappa_1, \kappa_2\}$. Let $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ be a distribution policy over $\mathcal{N}$ that distributes each fact onto every node in $\mathcal{N}$. Since every fact is present on both nodes, $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_1)$ and $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa_2)$ for every valuation V for $\mathcal{Q}$ over U. Assuming U is not empty, it trivially follows that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ holds, but $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ doesn't. ■

The construction of $\boldsymbol{P}$ in Example 5.3 whereby the complete instance is duplicated onto two nodes makes $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ true and $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ false for every query $\mathcal{Q}$, assuming there is at least one valuation V for $\mathcal{Q}$ over U. Since limiting ourself to only queries with no valuations is rather useless from a practical point of view, we conclude that a constraint on the considered distribution policies will be necessary to make $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ imply $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$.

We now consider the family of nonreplicating distribution policies $\mathcal{P}_{nrep}$. A distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ is in $\mathcal{P}_{nrep}$ if and only if it does not replicate any fact over U onto multiple nodes. More formally, a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over a network $\mathcal{N}$ is in $\mathcal{P}_{nrep}$ if and only if $rfacts_{\boldsymbol{P}}(\kappa_1) \cap rfacts_{\boldsymbol{P}}(\kappa_2) = \varnothing$ for every pair of nodes $\kappa_1, \kappa_2 \in \mathcal{N}$ with $\kappa_1 \neq \kappa_2$.

Observe that the distribution policy $\boldsymbol{P}$ used in Example 5.2 is a nonreplicating distribution policy. This example already illustrates that nonreplicating distribution policies on itself aren't sufficient to let $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ imply $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$. This desired result is achievable if nonreplicating distribution policies are combined with strongly minimal queries.

Proposition 5.4. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}[sm]$ be a strongly minimal conjunctive query with inequalities and let $\boldsymbol{P} \in \mathcal{P}_{nrep}$ be a nonreplicating distribution policy over a network $\mathcal{N}$. $PC_{set}(\boldsymbol{P}, \mathcal{Q})$ holds if and only if $PC_{bag}(\boldsymbol{P}, \mathcal{Q})$ holds.*

Proof. (if): This direction trivially follows from Proposition 5.1.

(only if): Let $\mathcal{Q}$ and $\boldsymbol{P}$ be as described in Proposition 5.4. Assume $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ holds. We prove that $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ holds as well.

Since $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$, there is a node $\kappa \in \mathcal{N}$ for every minimal valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Since $\mathcal{Q}$ is a strongly minimal conjunctive query, every valuation V for $\mathcal{Q}$ is minimal. As a result, there is at least one node $\kappa \in \mathcal{N}$ for every valuation V for $\mathcal{Q}$ over U such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

Notice furthermore that at the same time there is at most one node $\kappa \in \mathcal{N}$ for every valuation V for $\mathcal{Q}$ over U with this property. Indeed, if there would be two different nodes $\kappa_1, \kappa_2 \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_1)$ and $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$, then $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_1) \cap rfacts_{\boldsymbol{P}}(\kappa_2)$. But this would contradict our assumption that $rfacts_{\boldsymbol{P}}(\kappa_1) \cap rfacts_{\boldsymbol{P}}(\kappa_2) = \varnothing$, since $\boldsymbol{P}$ is a nonreplicating policy.

We conclude that for every valuation V for $\mathcal{Q}$ over U there is exactly one node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Thus, according to Proposition 4.6, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ holds. □

5.2 Transferability

When comparing the characterizations for transferability under set semantics (Proposition 3.12) and transferability under bag semantics (Proposition 4.29), there does not seem to be an immediate relation. The former is based on the covering of minimal valuations, whereas the latter is based on the observation that some facts are guaranteed to be grouped together on the same node. Furthermore, this notion of grouped facts is useless under set semantics as we can always isolate the required facts for a valuation on a separate node.

One might think that Proposition 5.1 would be sufficient to deduce the fact that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ implies $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ for every pair of conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$ in **CQ**. Unfortunately, this is not the case. In this section we explain why we can't directly derive from Proposition 5.1 that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ implies $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$. Furthermore, we prove by counterexample that, in general, $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ does not imply $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ and vice versa. Lastly, we show that a restriction to strongly minimal queries and nonreplicating distribution policies suffices to let $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ coincide.

We first focus on the direction from set semantics to bag semantics. Just like parallel-correctness, transferability under set semantics does not imply transferability under bag semantics when considering queries in **CQ**:

Example 5.5. For a counterexample showing that transferability under set semantics does not imply transferability under bag semantics, consider the following two conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$,

$$\mathcal{Q} : H() \leftarrow R(), S(),$$
$$\mathcal{Q}' : H() \leftarrow R().$$

Since $\mathcal{Q}$ uses no variables, there is exactly one valuation V_1 for $\mathcal{Q}$ which requires the facts $R()$ and $S()$. Analogously, there is exactly one valuation V_1' for $\mathcal{Q}'$, requiring only $R()$. It immediately follows that every valuation V' for $\mathcal{Q}'$ is covered by a valuation V for $\mathcal{Q}$. Therefore, $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ holds.

Let now $\boldsymbol{P}$ be a distribution policy over a network $\mathcal{N} = \{\kappa_1, \kappa_2\}$, assigning both $R()$ and $S()$ to κ_1 and only $R()$ to κ_2. $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ holds, as the required facts for V_1 only meet at κ_1. However, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q}')$ doesn't hold because the required fact for V_1' is present on both nodes. Therefore, parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ under bag semantics. We conclude that $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ does not imply $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$. ■

Next, we focus on the direction from bag semantics to set semantics. It might be tempting to believe that Proposition 5.1 combined with the definition of transferability directly implies that transferability under bag

semantics would imply transferability under set semantics. Unfortunately, this is not correct. Indeed, assume two conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$ in $\mathbf{CQ}^{\neq}$ such that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$. According to the definition of transferability and Proposition 5.1, we know that the following three statements are true for every distribution policy $\boldsymbol{P}$:

1. $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ implies $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q}')$,
2. $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ implies $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$, and
3. $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q}')$ implies $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$.

We cannot infer from these rules that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ implies $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$, as there might exist a distribution policy $\boldsymbol{P}'$ such that $PC_{\text{set}}(\boldsymbol{P}', \mathcal{Q})$ holds but $PC_{\text{set}}(\boldsymbol{P}', \mathcal{Q}')$, $PC_{\text{bag}}(\boldsymbol{P}', \mathcal{Q})$ and $PC_{\text{bag}}(\boldsymbol{P}', \mathcal{Q}')$ do not. Notice that this distribution policy $\boldsymbol{P}'$ is a counterexample of $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ while satisfying all three conditions stated above.

Although the possible existence of such a distribution policy $\boldsymbol{P}'$ does not interfere with Proposition 5.1 and the definition of transferability, we still need to show that such a policy actually exists to prove that transferability under bag semantics does not imply transferability under set semantics. In the following counterexample, we construct such a policy for two conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$ satisfying $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$.

Example 5.6. Consider the following two conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$,

$$\mathcal{Q} : H(x, y, z) \leftarrow R(x, y), R(y, z).$$
$$\mathcal{Q}' : H(w, x, y, z) \leftarrow R(w, x), R(x, y), R(y, z).$$

Notice that these queries are the same as in Example 4.15, apart from the head atoms. In Example 4.15 we already motivated that every distribution policy $\boldsymbol{P}$ should distribute the facts in such a way that every valuation is only satisfiable on the same node κ. Furthermore, we showed that parallel-correctness therefore transfers from $\mathcal{Q}$ to $\mathcal{Q}'$. This reasoning is still applicable, as the changed head atoms are not important during the reasoning. We conclude that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ holds.[1]

Next, we show that $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ doesn't hold. Let $U = \{1, \ldots, k\}$ be a finite universe. Assume a distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over a network $\mathcal{N}$, distributing every pair of facts over the relation R onto a different node in $\mathcal{N}$. Since U is finite, a finite number of nodes in $\mathcal{N}$ is sufficient for this construction.

By construction of $\boldsymbol{P}$ there trivially is a node $\kappa \in \mathcal{N}$ for every valuation V for $\mathcal{Q}$ over U with $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. It immediately follows that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$. Therefore, it now suffices to show that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ doesn't

[1] Alternatively, we could have applied Proposition 4.29 to obtain the same result.

hold to prove that parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ under set semantics.

Since every variable in $\mathcal{Q}'$ appears in the head atom, every valuation for $\mathcal{Q}'$ is a minimal valuation. Consider for example the valuation $V' = \{w \mapsto 0, x \mapsto 1, y \mapsto 2, z \mapsto 3\}$ for $\mathcal{Q}'$ over U. This valuation V' derives the fact $V(head_{\mathcal{Q}'}) = H(0,1,2,3)$, thereby requiring the facts $V(body_{\mathcal{Q}'}) = \{R(0,1), R(1,2), R(2,3)\}$. It can easily be seen that there is no other valuation for $\mathcal{Q}'$ over U deriving $H(0,1,2,3)$ and requiring only a strict subset of the facts in $V(body_{\mathcal{Q}'})$, so V' is a minimal valuation for $\mathcal{Q}'$. By Proposition 3.8, $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ can only hold if there is a node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Notice however that $V(body_{\mathcal{Q}})$ contains three facts, whereas every node in $\mathcal{N}$ contains by construction of $\boldsymbol{P}$ only two facts. Therefore, $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ cannot hold, and we conclude that parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ under set semantics.

Alternatively, we could apply Proposition 3.12 directly. This property states that every minimal valuation for $\mathcal{Q}'$ is covered by a valuation V for $\mathcal{Q}$ if $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ holds. More specifically, since V' is a minimal valuation for $\mathcal{Q}'$, it should be covered by a valuation V for $\mathcal{Q}$. However, this is not the case, as every valuation V for $\mathcal{Q}$ clearly requires at most two facts, whereas V' requires three facts. This alternative approach confirms that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ does not hold.

Since parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ under bag semantics but not under set semantics, these conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$ prove that in general transferability under bag semantics does not imply transferability under set semantics. ■

Just as for parallel-correctness, a restriction on the considered queries and distribution policies might be useful to let transferability under set and bag semantics coincide. According to Definition 3.3, transferability is however defined over *every* distribution policy. We therefore provide a slightly modified definition of transferability taking into account restricted sets of distribution policies.

Definition 5.7. For two queries $\mathcal{Q}$ and $\mathcal{Q}'$ over the same input schema, *parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ relative to a set of distribution policies $\mathcal{P}$* if $\mathcal{Q}'$ is parallel-correct under every distribution policy in $\mathcal{P}$ for which $\mathcal{Q}$ is parallel-correct.

We now describe a restriction on both conjunctive queries and distribution policies to let transferability under set semantics coincide with transferability under bag semantics.

Proposition 5.8. *Let $\mathcal{Q}, \mathcal{Q}' \in \mathbf{CQ}^{\neq}[sm]$ be strongly minimal conjunctive queries with inequalities. $PCT_{set}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ holds if and only if $PCT_{bag}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ holds.*

Proof. (if) Assume $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ holds. We show that $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ holds. Towards a contradiction, assume $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ does not hold, witnessed by a distribution policy $\boldsymbol{P} \in \mathcal{P}_{nrep}$. In other words, $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ does hold, but $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ does not. According to Proposition 5.4, the results under bag semantics have to be the same. That is, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ does hold and $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q}')$ does not. But this observation contradicts with our initial assumption that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$. We therefore conclude that $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$.

(only if) This part of the proof is analogous to the previous part. Assume $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ while $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ does not hold. There has to be a distribution policy $\boldsymbol{P} \in \mathcal{P}_{nrep}$ such that $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ holds, but $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q}')$ doesn't, meaning that $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ does hold, but $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q}')$ does not. The latter however contradicts our initial assumption, indicating that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$ holds if $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ relative to $\mathcal{P}_{nrep}$. □

Notice that strongly minimal conjunctive queries and nonreplicating distribution policies aren't a necessary condition to let transferability under set and bag semantics coincide for a pair of conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$.

Example 5.9. To illustrate that transferability under bag and set semantics might coincide for arbitrary queries in $\mathbf{CQ}^{\neq}$ and arbitrary distribution policies, consider the case where $\mathcal{Q}$ equals $\mathcal{Q}'$. In this case, parallel-correctness trivially transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ (and vice versa) under both set and bag semantics, implying that $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ coincide. ■

Chapter 6

Modifying the distributed evaluation model

The characterization for parallel-correctness under bag semantics described in Proposition 4.6 constrains the possible distribution policies for a given conjunctive query. Depending on the conjunctive query $\mathcal{Q}$, this condition might even imply that every distribution policy for $\mathcal{Q}$ takes little to no advantage of the distributed environment. We already saw an extreme case of such a conjunctive query $\mathcal{Q}$ in Example 4.15. The conjunctive query $\mathcal{Q}$ in this example required every valuation to be evaluated on the same node in the network. It can easily be seen that such a policy has little practical use, as it cannot utilize the different nodes in the network to distribute the work.

In this chapter, we modify the definition of the distributed evaluation model, allowing a relaxed condition for parallel-correctness under bag semantics. This modification allows different nodes to be responsible for all the facts required by some valuation V. If multiple such nodes for V exist, exactly one node will eventually derive a fact based on V.

6.1 Definitions

Ordered networks An ordered network $\mathcal{N}$ is a nonempty finite *ordered* sequence $(\kappa_1, \ldots, \kappa_n)$ of nodes. Let κ_i and κ_j be two nodes in an ordered network $\mathcal{N}$. We say that κ_i *precedes* κ_j, denoted $\kappa_i <_{\mathcal{N}} \kappa_j$, if κ_i occurs in $\mathcal{N}$ before κ_j.

Notice that an ordered network $\mathcal{N}$ can be seen as a conventional network defined in Chapter 2 with the additional function $<_{\mathcal{N}}$ defining a total order over the nodes in $\mathcal{N}$. Therefore, the definitions of distribution policies and local instances mentioned in Chapter 2 are applicable to ordered networks as well. Let $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ be a distribution policy. A node $\kappa \in \mathcal{N}$ is *responsible* for a valuation V for a query $\mathcal{Q}$ over U if $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$

and there is no node $\kappa' \in \mathcal{N}$ with $\kappa' <_{\mathcal{N}} \kappa$ and $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa')$. Intuitively, the node responsible for a valuation V is the first node in the ordered network satisfying the required facts for V.

Conventional distributed evaluation Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over an ordered network $\mathcal{N}$. This ordered network $\mathcal{N}$ can be seen as a conventional network defined in Chapter 2 as well. Therefore, the conventional definitions for the distributed evaluation of a query $\mathcal{Q}$ on instance I under set and bag semantics, denoted respectively $[\mathcal{Q}, \boldsymbol{P}]_{\text{set}}(I)$ and $[\mathcal{Q}, \boldsymbol{P}]_{\text{bag}}(I)$, are still applicable.

Distributed evaluation under set semantics Let $\boldsymbol{P}$ be a distribution policy over an ordered network $\mathcal{N}$. The result of the distributed evaluation of a query $\mathcal{Q}$ on instance I on a node $\kappa \in \mathcal{N}$ under set semantics, denoted $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{set}}(I, \kappa)$, is defined as

$$[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{set}}(I, \kappa) = \bigcup_{V \in \mathcal{V}_\kappa} [\mathcal{Q}, V]_{\text{set}}(\textit{loc-inst}_{\text{bag}, \boldsymbol{P}, I}(\kappa))$$

with $\mathcal{V}_\kappa$ the set of valuations for $\mathcal{Q}$ having κ as responsible node. The result of the distributed evaluation of a query $\mathcal{Q}$ on instance I under set semantics, denoted $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{set}}(I)$ is defined as

$$[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{set}}(I) = \bigcup_{\kappa \in \mathcal{N}} [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{set}}(I, \kappa).$$

Intuitively, each node in an ordered network only considers the valuations it is responsible for while evaluating $\mathcal{Q}$, after which the set union over all these results is taken to produce the final result.

Distributed evaluation under bag semantics Let $\boldsymbol{P}$ be a distribution policy over an ordered network $\mathcal{N}$. The result of the distributed evaluation of a query $\mathcal{Q}$ on instance I on a node $\kappa \in \mathcal{N}$ under bag semantics, denoted $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{bag}}(I, \kappa)$, is defined as

$$[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{bag}}(I, \kappa) = \bigcup_{V \in \mathcal{V}_\kappa} {}_B [\mathcal{Q}, V]_{\text{bag}}(\textit{loc-inst}_{\text{bag}, \boldsymbol{P}, I}(\kappa))$$

with $\mathcal{V}_\kappa$ the set of valuations for $\mathcal{Q}$ having κ as responsible node. The result of the distributed evaluation of a query $\mathcal{Q}$ on instance I under bag semantics, denoted $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{bag}}(I)$ is defined as

$$[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{bag}}(I) = \bigcup_{\kappa \in \mathcal{N}} {}_B [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}]_{\text{bag}}(I, \kappa).$$

Analogously to set semantics, each node in an ordered network intuitively only considers the valuations it is responsible for while evaluating $\mathcal{Q}$, after which the bag union over all these results is taken to produce the final result.

Parallel-correctness under ordered networks We slightly modify the existing definitions of parallel-correctness (Definition 3.1 and Definition 3.2) for these distributed evaluations over ordered networks.

A query $\mathcal{Q}$ is *parallel-correct on an instance I under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$* if $\mathcal{Q}(I) = [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$. When this definition is lifted to all instances, we get the following definition for parallel-correctness under ordered networks:

Definition 6.1. A query $\mathcal{Q}$ is *parallel-correct under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$* if $\mathcal{Q}$ is parallel-correct on all instances I under $\boldsymbol{P}$ over $\mathcal{N}$.

6.2 The modified model as a single-round MPC model

Under the modified model, a node only uses a valuation to derive a fact if no preceding node in the ordered network uses this valuation. A naive approach would be to let each node communicate with its preceding nodes in the ordered network to check if they already applied the valuation. This naive approach not only causes a massive communication overhead on larger instances and networks, it does furthermore no longer fit in the single-round MPC model as it requires a second round of communication.

These problems can be avoided if we assume that each node in the ordered network has knowledge of both the applied distribution policy and the ordered network itself. With this knowledge, a node in the ordered network can perform the necessary checks during the computation phase, since all the necessary information is locally available. This approach no longer requires a second communication round and consequently fits in the single-round MPC model.

A second remark on the modified model is that it tends to skew the number of produced results over the different nodes. Consider for example the extreme case where every fact is distributed over every node. It can easily be seen that in this case only the first node produces results because it is the responsible node for every valuation. In practise, this is however not necessarily a problem, as the aim of distribution policies is to distribute the work over the different nodes instead of replicating it. In other words, a practical distribution policy aims to distribute the data as much as possible, giving each node only a small part of the global data. Consequently, most of the valuations will only be satisfiable on one or a couple of nodes, meaning that the amount of skew will often be relatively small in practical scenarios.

6.3 The modified model under set semantics

Assume a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$, a distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ and an instance I. If we compare the definitions of $[\mathcal{Q}, \boldsymbol{P}](I)$ and $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$, we can intuitively deduce that both definitions produce the same result. The main difference is that the distributed evaluation based on the modified model explicitly removes duplicates by considering each valuation at most once, whereas the standard distributed evaluation implicitly removes these duplicates while taking the set union over the results on the different nodes.

Proposition 6.2. *For each query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$, for each distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ and for each instance I it holds that $[\mathcal{Q}, \boldsymbol{P}](I)$ equals $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$.*

Proof. Assume a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$, a distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ and an instance I. We prove that $f \in [\mathcal{Q}, \boldsymbol{P}](I)$ if and only if $f \in [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ for every fact f over U.

(if) Assume $f \in [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$. By definition, there exist a valuation V for $\mathcal{Q}$ over U and a node $\kappa \in \mathcal{N}$ responsible for V with $V(\mathit{head}_{\mathcal{Q}}) = f$ and $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{loc\text{-}inst}_{\mathrm{set},\boldsymbol{P},I}(\kappa)$. Since there is a node κ in $\mathcal{N}$ containing all the required facts for V, it clearly follows by definition of $[\mathcal{Q}, \boldsymbol{P}](I)$ that $f \in [\mathcal{Q}, \boldsymbol{P}](I)$.

(only if) Assume $f \in [\mathcal{Q}, \boldsymbol{P}](I)$. By definition, there exist a valuation V for $\mathcal{Q}$ over U and a node κ_V in $\mathcal{N}$ with $V(\mathit{head}_{\mathcal{Q}}) = f$ and $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{loc\text{-}inst}_{\mathrm{set},\boldsymbol{P},I}(\kappa_V)$. Let $N \subseteq \mathcal{N}$ denote the subset of nodes κ in $\mathcal{N}$ having $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. By definition of an ordered network, there is a node $\kappa \in N$ preceding all the other nodes in N. This node κ is responsible for V. It can easily be seen by definition of $\mathit{rfacts}_{\boldsymbol{P}}$ and $\mathit{loc\text{-}inst}_{\mathrm{set},\boldsymbol{P},I}$ that $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{loc\text{-}inst}_{\mathrm{set},\boldsymbol{P},I}(\kappa)$ if $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{loc\text{-}inst}_{\mathrm{set},\boldsymbol{P},I}(\kappa_V)$.

Since all the required facts for V are in the local instance on κ and since κ is responsible for V, this node κ will derive the fact $f = V(\mathit{head}_{\mathcal{Q}})$. We conclude that $f \in [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$. □

Since the result under the conventional distributed evaluation and the result under the distributed evaluation over an ordered network are always the same, we conclude that a characterization for the conventional definition parallel-correctness under set semantics is a characterization for parallel-correctness under an ordered network as well. We slightly modify Condition 3.7 and Proposition 3.8 to include ordered networks:

Condition 6.3. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a query and $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ a distribution policy over a ordered network $\mathcal{N}$. For every minimal valuation V for $\mathcal{Q}$ over U, there is a node $\kappa \in \mathcal{N}$ such that $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$.*

Proposition 6.4. *Under set semantics, a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is parallel-correct under distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ if and only if Condition 6.3 is satisfied.*

Since the characterization for parallel-correctness didn't change under this modified model, the characterization for transferability, described in Proposition 3.12, is applicable to ordered networks as well.

Proposition 6.5. *For queries $\mathcal{Q}, \mathcal{Q}' \in \mathbf{CQ}^{\neq}$, parallel-correctness over ordered networks transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if $\mathcal{Q}$ covers $\mathcal{Q}'$.*

Recall from Definition 3.11 that a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ covers a query $\mathcal{Q}' \in \mathbf{CQ}^{\neq}$ if and only if for every minimal valuation V' for $\mathcal{Q}'$ there is a minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$.

6.4 Parallel-correctness under bag semantics

Unlike set semantics, The result $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ of a distributed evaluation based on an ordered network under bag semantics is not always the same as the result $[\mathcal{Q}, \boldsymbol{P}](I)$ based on the conventional approach. Assume for example a valuation V for $\mathcal{Q}$ for which there are multiple nodes in $\mathcal{N}$ responsible for all the required facts of V. If the required facts for V are in the considered instance I, the fact $f = V(head_{\mathcal{Q}})$ will be derived multiple times in the conventional model, whereas it will only be derived by the node responsible for V in our modified model. Therefore, the multiplicity of f will be strictly higher in $[\mathcal{Q}, \boldsymbol{P}](I)$ than in $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$.

Proposition 6.6. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a query and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over an ordered network $\mathcal{N}$. For every instance I over U, it holds that $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I) \subseteq [\mathcal{Q}, \boldsymbol{P}](I)$ and $facts([\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) = facts([\mathcal{Q}, \boldsymbol{P}](I))$.*

Proof. Assume a query $\mathcal{Q}$, distribution policy $\boldsymbol{P}$, ordered network $\mathcal{N}$ and instance I as in Proposition 6.6. It trivially follows from Proposition 6.2 that $facts([\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$ equals $facts([\mathcal{Q}, \boldsymbol{P}](I))$. It now suffices to show that $mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) \leq mul(f, [\mathcal{Q}, \boldsymbol{P}](I))$ for every fact f appearing in $facts([\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$ to prove that $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I) \subseteq [\mathcal{Q}, \boldsymbol{P}](I)$.

To this end, let f be a fact in $facts([\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$. By definition, the multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ is determined by all the valuations V deriving f on the node responsible for V. The multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}](I)$ on the other hand is determined by all the valuations V deriving f on some node (even if this node is not the node responsible for V in the ordered network). We can easily see that every valuation V on a node κ contributing to the multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ equally contributes to the multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}](I)$, but the converse is not true in general. We conclude that $mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) \leq mul(f, [\mathcal{Q}, \boldsymbol{P}](I))$. □

There is another interesting difference between the conventional model and the modified model: Under the modified model, a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is always parallel-sound under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$. Recall that a query $\mathcal{Q}$ is parallel-sound under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$ if $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I) \subseteq \mathcal{Q}(I)$ for every instance I.

Proposition 6.7. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a query and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over an ordered network $\mathcal{N}$. For every instance I over U, it holds that $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I) \subseteq \mathcal{Q}(I)$.*

Proof. Assume a query $\mathcal{Q}$, distribution policy $\boldsymbol{P}$, ordered network $\mathcal{N}$ and instance I. We prove that $mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) \leq mul(f, \mathcal{Q}(I))$ for every fact $f \in facts([\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$.

Recall that by definition

$$mul(f, \mathcal{Q}(I)) = \sum_{S \in \mathcal{V}} mul(f, [\mathcal{Q}, S](I)) \tag{1}$$

with $\mathcal{V}$ the set of satisfying valuations for $\mathcal{Q}$ on I deriving f. Furthermore,

$$mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) = \sum_{\kappa \in \mathcal{N}} \sum_{T \in \mathcal{V}_\kappa} mul(f, [\mathcal{Q}, T](loc\text{-}inst_{\boldsymbol{P},I}(\kappa))) \tag{2}$$

with $\mathcal{V}_\kappa$ the set of satisfying valuations for $\mathcal{Q}$ on $loc\text{-}inst_{\boldsymbol{P},I}(\kappa)$ deriving f and having κ as responsible node.

Since every term appearing in the second equation is based on a satisfied valuation T when evaluated over a local instance, this valuation T is satisfied on the global instance I as well. Therefore, this term based on T will appear in the first equation as well. Notice furthermore that every valuation V used in the first equation can appear at most once in a term in the second valuation. Indeed, if V would appear in two different terms, V would derive f on multiple nodes. But this contradicts with our definition stating that every valuation V only derives facts on the single node responsible for it.

We conclude that for every term appearing in the second equation, there is an equivalent term in the first equation. The opposite is however not necessarily true, implying that $mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)) \leq mul(f, \mathcal{Q}(I))$. □

Notice that, in general, a conjunctive query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is not parallel-complete under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$. In other words, $\mathcal{Q}(I) \subseteq [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ does not necessarily hold. A trivial counterexample is a distribution policy that does not assign facts to nodes, thereby always producing the empty result.

These differences between the modified and the conventional distributed evaluation model intuitively lead to a more relaxed condition for parallel-correctness under bag semantics.

Recall that Condition 4.3 is both necessary and sufficient for parallel-correctness under bag semantics when we are considering the conventional distributed evaluation model. This condition states that every valuation for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ has to be satisfiable on exactly one node in the network. During the proof of Lemma 4.4 we show that the required facts for a valuation V cannot meet at more than one node, as the resulting multiplicity of the derived fact would be too high in the result.

Under the modified distributed evaluation model, however, the multiplicity of a fact cannot be too high, as a query in $\mathbf{CQ}^{\neq}$ is always parallel-sound under every distribution policy according to Proposition 6.7. Therefore, the required facts for a valuation can meet at more than one node, implying that Condition 4.3 is no longer necessary for parallel-correctness under bag semantics.

Based on these observations, we formulate a condition closely related to Condition 4.3 that is both necessary and sufficient for parallel-correctness over ordered networks under bag semantics.

Condition 6.8. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ be a conjunctive query with inequalities and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over an ordered network $\mathcal{N}$. For every valuation V for $\mathcal{Q}$ over U, there is at least one node $\kappa \in \mathcal{N}$ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Before proving that Condition 6.8 is a necessary and sufficient condition for parallel-correctness over ordered networks, we first consider the following lemma, closely related to Lemma 4.5.

Lemma 6.9. *A query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is not parallel-correct under a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ if there exists a valuation V for $\mathcal{Q}$ with no node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proof sketch. The proof is analogous to the proof of Lemma 4.5. If there is a valuation V that derives a fact f on the global instance I, but not on the local instance of a node in the network, the multiplicity of f will be lower in $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$ than in $\mathcal{Q}(I)$, unless there is some kind of compensation.

During the proof of Lemma 4.5, it is shown that the only possible compensation is another valuation V' that derives f on multiple nodes. Under the modified distributed evaluation model, this construction is by definition impossible because V' only derives f on the node $\kappa \in \mathcal{N}$ that is responsible for V'. We conclude that it is impossible to compensate for the multiplicity of f being too low. Therefore, $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$ over the ordered network $\mathcal{N}$. □

We are now ready to prove that Condition 6.8 is both necessary and sufficient for parallel-correctness under the modified distributed evaluation model.

Proposition 6.10. *Under bag semantics, a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ is parallel-correct under distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ if and only if Condition 6.8 is satisfied.*

Proof. (if) Assume Condition 6.8 is satisfied. We prove that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ over $\mathcal{N}$. According to Proposition 6.7, parallel-soundness is already guaranteed, so we only need to prove parallel-completeness. In other words, $mul(f, \mathcal{Q}(I)) \leq mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$ has to hold for every fact $f \in facts(\mathcal{Q}(I))$.

By definition, the multiplicity of f in $\mathcal{Q}(I)$ is determined by the set of valuations $\mathcal{V}$ deriving f on instance I. Since the required facts for every valuation $V \in \mathcal{V}$ meet on at least one node in the network by assumption, we conclude that every valuation $V \in \mathcal{V}$ contributing to the final multiplicity of f in $\mathcal{Q}(I)$ contributes an equal amount to the multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I)$. It immediately follows that $mul(f, \mathcal{Q}(I)) \leq mul(f, [\mathcal{Q}, \boldsymbol{P}, \mathcal{N}](I))$.

(only if) The proof is by contraposition. Assume Condition 6.8 is not satisfied. In other words, there exists a valuation V for $\mathcal{Q}$ with no node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. It immediately follows from Lemma 6.9 that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$ over $\mathcal{N}$. □

6.5 Complexity of parallel-correctness under bag semantics

Although the characterization for parallel-correctness under bag semantics over an ordered network is less strict, there is no immediate change in the upper bounds for deciding parallel-correctness. We first reformulate the problem to include ordered networks.

	$\mathbf{PC}_{\text{ord}}(\mathcal{C}, \mathcal{P})$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$
Question:	Is $\mathcal{Q}$ parallel-correct under $\boldsymbol{P}$ over an ordered network?

Proposition 6.11. *The problem $\mathbf{PC}_{ord}(\mathcal{C}, \mathcal{P})$ is in Π_2^p for every query class $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ and every distribution policy class $\mathcal{P} \in \{\mathcal{P}_{fin}\} \cup \mathfrak{P}_{nondet}$.*

Proof. Let k be fixed and let $\langle \mathcal{Q}, \boldsymbol{P} \rangle$ be an input for $\mathbf{PC}_{\text{ord}}(\mathbf{CQ}^{\neq}, \mathcal{P}^k_{nondet})$, with $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ represented by a tuple $(n, \mathcal{A}_{\boldsymbol{P}})$. According to Proposition 6.10, it suffices to show that there is a Π_2^p-algorithm that checks whether for each valuation V for $\mathcal{Q}$ over U there is a node κ such that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.

As explained in the proof of Proposition 4.7, there exists a verifier M deciding whether $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ in polynomial time on $V(\mathit{body}_\mathcal{Q})$ and κ as input. We use this verifier M to construct a Π_2^p-algorithm deciding $\mathbf{PC}_{\text{ord}}(\mathbf{CQ}^{\neq}, \mathcal{P}^k_{\mathit{nondet}})$.

The algorithm accepts if and only if for every valuation V for Q over U, there is a node κ and a certificate c such that the verifier M accepts on input $\langle V(\mathit{body}_\mathcal{Q}), \kappa\rangle$ with certificate c. Since deciding whether the verifier M accepts on input $\langle V(\mathit{body}_\mathcal{Q}), \kappa\rangle$ with certificate c is obviously possible in polynomial time, it can easily be seen that this algorithm is indeed in Π_2^p.

Note that this result also holds for query class $\mathbf{CQ}$ and policy class $\mathcal{P}_{\mathit{fin}}$, since $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$ and $\mathcal{P}_{\mathit{fin}} \subseteq \mathcal{P}^2_{\mathit{nondet}}$. □

Again, we can lower this upper bound by restricting ourselves to distribution policies in $\mathfrak{P}_{\mathit{det}}$, thereby dropping the used existential quantifier.

Proposition 6.12. *The problem* $\mathbf{PC}_{\mathit{ord}}(\mathcal{C}, \mathcal{P})$ *is in* coNP *for every query class* $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ *and every distribution policy class* $\mathcal{P} \in \mathfrak{P}_{\mathit{det}} \cup \{\mathcal{P}_{\mathit{fin}}\}$.

Proof. It suffices to show that the complement of the problem $\mathbf{PC}_{\text{ord}}(\mathcal{C}, \mathcal{P})$, denoted $\overline{\mathbf{PC}_{\text{ord}}(\mathcal{C}, \mathcal{P})}$, is in NP for every query class $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ and every distribution policy class $\mathcal{P} \in \mathfrak{P}_{\mathit{det}} \cup \{\mathcal{P}_{\mathit{fin}}\}$.

	$\overline{\mathbf{PC}_{\text{ord}}(\mathcal{C}, \mathcal{P})}$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	distribution policy $\boldsymbol{P} \in \mathcal{P}$
Question:	Is $\mathcal{Q}$ not parallel-correct under $\boldsymbol{P}$
	over an ordered network?

Let k be fixed. We construct a nondeterministic algorithm deciding $\overline{\mathbf{PC}_{\text{ord}}(\mathbf{CQ}^{\neq}, \mathcal{P}^k_{\mathit{det}})}$ on input $\langle \mathcal{Q}, \boldsymbol{P}\rangle$ with $\mathcal{Q} \in \mathbf{CQ}^{\neq}$ and $\boldsymbol{P} \in \mathcal{P}^k_{\mathit{det}}$. By assumption, $\boldsymbol{P}$ is representable by a tuple $(\mathcal{N}, n, \mathcal{A}_P)$. According to Proposition 6.10, $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$ if and only if there is a valuation V for $\mathcal{Q}$ over U such that there is no node κ with $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. We use this property to construct the nondeterministic algorithm as follows: Guess a valuation V and check if there is a node $\kappa \in \mathcal{N}$ having $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. The algorithm rejects if this is true. Otherwise, it accepts.

Since valuations are mappings of variables appearing in $\mathcal{Q}$ to values in U and since values in U can be represented by a string of length n or less, it is possible to guess a valuation V in polynomial time. Furthermore, checking whether there is a node $\kappa \in \mathcal{N}$ with $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ is also possible in polynomial time, as the nodes are an explicit part of the input and $\mathcal{A}_P$ can be used to decide $V(\mathit{body}_\mathcal{Q}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$ in polynomial time for each node $\kappa \in \mathcal{N}$. Thus, the nondeterministic algorithm described above runs in polynomial time.

Notice that this reasoning holds for conjunctive queries in **CQ** as well, since $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$. □

Unfortunately, the construction in the lower bound proof of Proposition 4.9 isn't directly applicable to parallel-correctness under ordered networks, as it uses the fact that a query isn't parallel-correct under a distribution policy over a regular network if a valuation is satisfiable on two nodes.

6.6 Transferability under bag semantics

The characterization for transferability under bag semantics when considering regular networks, described in Proposition 4.29, is based on the observation that some facts always group together on the same node. Under the modified model however, this notion of *impFacts* is no longer useful, as this grouping of facts is no longer required. Indeed, a distribution policy under which a conjunctive query $\mathcal{Q}$ is parallel-correct can always map the required facts for each valuation for $\mathcal{Q}$ onto a different node, implying that $\mathit{impFacts}(V, \mathcal{Q})$ would always equal $V(\mathit{body}_{\mathcal{Q}})$ under this modified model.

Condition 6.13. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. For each valuation V' for $\mathcal{Q}'$ over a universe U, there is a valuation V for $\mathcal{Q}$ over U such that $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq V(\mathit{body}_{\mathcal{Q}})$.*

Condition 6.13 intuitively is a sufficient condition for parallel-correctness transferring from $\mathcal{Q}$ to $\mathcal{Q}'$ over ordered networks. We show furthermore that Condition 6.13 is a necessary condition as well.

Proposition 6.14. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in $\mathbf{CQ}^{\neq}$. Parallel-correctness over ordered networks transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if Condition 6.13 is satisfied.*

Proof. (if) Assume Condition 6.13 is satisfied. Consider an arbitrary distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over an ordered network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct. We show that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well by proving that for each valuation V' for Q' there is a node $\kappa \in \mathcal{N}$ with $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. To this end, consider an arbitrary valuation V' for $\mathcal{Q}'$. By assumption, there is a valuation V for $\mathcal{Q}$ such that $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq V(\mathit{body}_{\mathcal{Q}})$. According to Proposition 6.10, there exists a node $\kappa \in \mathcal{N}$ with $V(\mathit{body}_{\mathcal{Q}}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. It immediately follows that $V'(\mathit{body}_{\mathcal{Q}'}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$. We conclude that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ over the ordered network $\mathcal{N}$.

(only if) The proof is by contraposition. Assume Condition 6.13 isn't satisfied. We show that parallel-correctness under ordered networks doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ by constructing a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct, but $\mathcal{Q}'$ isn't.

By assumption, there is a valuation V' for $\mathcal{Q}'$ over a universe U such that there is no valuation V for $\mathcal{Q}$ over U with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$. Let $\mathcal{D}$ be a schema containing all relations used in $\mathcal{Q}$ and $\mathcal{Q}'$ and let $V'(body_{\mathcal{Q}'}) = \{f_1, \ldots, f_n\}$. The construction of a distribution policy $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ over an ordered network $\mathcal{N} = (\kappa_1, \ldots, \kappa_n)$ is as follows. For each node $\kappa_i \in \mathcal{N}$, we define $rfacts_{\boldsymbol{P}}(\kappa_i) = facts(\mathcal{D}, U) \setminus \{f_i\}$.

Notice that there is no node $\kappa \in \mathcal{N}$, with $V'(body_{\mathcal{Q}'}) \in rfacts_{\boldsymbol{P}}(\kappa)$. Therefore, according to Proposition 6.10, $\mathcal{Q}'$ isn't parallel-correct under $\boldsymbol{P}$ over the ordered network $\mathcal{N}$. It next suffices to show that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ over the ordered network $\mathcal{N}$ to conclude that parallel-correctness over ordered networks doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$. To this end, consider an arbitrary valuation V for $\mathcal{Q}$ over U. By assumption, $V'(body_{\mathcal{Q}'}) \not\subseteq V(body_{\mathcal{Q}})$. Therefore, there exists a fact $f_i \in V'(body_{\mathcal{Q}'})$ with $f_i \notin V(body_{\mathcal{Q}})$. By construction of $\boldsymbol{P}$, it now follows that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_i)$. We conclude that each valuation V for $\mathcal{Q}$ is satisfiable on at least one node $\kappa \in \mathcal{N}$. The query $\mathcal{Q}$ therefore satisfies Condition 6.8, implying that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ over the ordered network $\mathcal{N}$. □

6.7 Complexity of transferability under bag semantics

The high upper bound on the complexity of deciding transferability over regular networks given in Proposition 4.31 is mainly a consequence of the computation of *impFacts*. Over ordered networks, this computation is no longer necessary, resulting in an improved upper bound.

Before describing this upper bound, we first reformulate the problem of transferability to include ordered networks.

	$\mathbf{PC\text{-}Trans}_{\text{ord}}(\mathcal{C}, \mathcal{C}')$
Input:	Query $\mathcal{Q} \in \mathcal{C}$,
	query $\mathcal{Q}' \in \mathcal{C}'$
Question:	Does parallel-correctness over ordered networks
	transfer from $\mathcal{Q}$ to $\mathcal{Q}'$?

Proposition 6.15. *The problem* $\mathbf{PC\text{-}Trans}_{ord}(\mathcal{C}, \mathcal{C}')$ *is in* Π_2^p *for every query class* $\mathcal{C} \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$ *and every query class* $\mathcal{C}' \in \{\mathbf{CQ}, \mathbf{CQ}^{\neq}\}$.

Proof. Let $\langle \mathcal{Q}, \mathcal{Q}' \rangle$ be an input for $\mathbf{PC\text{-}Trans}_{\text{ord}}(\mathbf{CQ}^{\neq}, \mathbf{CQ}^{\neq})$. We construct an algorithm deciding transferability by returning true if and only if for every valuation V' for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$. Notice that this algorithm clearly is correct as it is a direct application of Condition 6.13.

Since deciding whether $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$ is obviously possible in polynomial time, the described algorithm is a Π_2^p-algorithm deciding $\mathbf{PC\text{-}Trans}_{\text{ord}}(\mathbf{CQ}^{\neq}, \mathbf{CQ}^{\neq})$. This result holds for queries in $\mathbf{CQ}$ as well, as $\mathbf{CQ} \subseteq \mathbf{CQ}^{\neq}$. □

6.8 Relation between set and bag semantics

Chapter 5 describes the relation of parallel-correctness and transferability between set and bag semantics. In this section we study this relation while considering ordered networks instead of regular networks. To facilitate our reasoning, we continue using our notations $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$, $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$, $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ to indicate parallel-correctness and transferability under set and bag semantics over regular networks. We furthermore use $PC_{\text{ord,set}}(\boldsymbol{P}, \mathcal{Q})$ and $PC_{\text{ord,bag}}(\boldsymbol{P}, \mathcal{Q})$ to denote the fact that the query $\mathcal{Q}$ is parallel-correct under distribution policy $\boldsymbol{P}$ over an ordered network under respectively set and bag semantics. Analogously, we use $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ to indicate that parallel-correctness over ordered networks transfers from a query $\mathcal{Q}$ to a query $\mathcal{Q}'$ under respectively set and bag semantics.

6.8.1 Parallel-Correctness

Recall from Proposition 5.1 that $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ always implies $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$. This remains the case under our modified model.

Proposition 6.16. *Let $\mathcal{Q}$ be a query in $\mathbf{CQ}^{\neq}$ and let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over an ordered network $\mathcal{N}$. If $PC_{ord,bag}(\boldsymbol{P}, \mathcal{Q})$, then $PC_{ord,set}(\boldsymbol{P}, \mathcal{Q})$.*

Proof. Let the query Q, policy $\boldsymbol{P}$ and ordered network $\mathcal{N}$ be as described in the proposition. Assume $PC_{\text{ord,bag}}(\boldsymbol{P}, \mathcal{Q})$ holds. By Proposition 6.10, for every valuation V for $\mathcal{Q}$ over U, there is a node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. It clearly follows that Condition 6.3 is satisfied, implying that $PC_{\text{ord,set}}(\boldsymbol{P}, \mathcal{Q})$ holds. □

On the other hand, parallel-correctness under set semantics doesn't necessarily imply parallel-correctness under bag semantics while considering ordered networks.

Example 6.17. Based on Example 5.2, Consider the conjunctive query $\mathcal{Q}$,

$$T(x) \leftarrow R(x), R(y)$$

and the ordered network $\mathcal{N} = (\kappa_1, \kappa_2)$. Assume a binary universe $U = \{a, b\}$. Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over $\mathcal{N}$ with $rfacts_{\boldsymbol{P}}(\kappa_1) = \{R(a)\}$ and $rfacts_{\boldsymbol{P}}(\kappa_2) = \{R(b)\}$.

As already explained in Example 5.2, the required facts for each minimal valuation meet on a node in the network, implying $PC_{\text{ord,set}}(\boldsymbol{P}, \mathcal{Q})$. There exists however a valuation $V = \{x \mapsto a, y \mapsto b\}$ for which there is no node $\kappa \in \mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Consequently, $PC_{\text{ord,bag}}(\boldsymbol{P}, \mathcal{Q})$ doesn't hold. ■

Restricting ourselves to strongly minimal queries solves the issue described in Example 6.17. Recall from Chapter 5 that we combined strongly minimal queries with nonreplicating distribution policies to let $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ coincide with $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$. Over ordered networks, the latter restriction is however no longer necessary.

Proposition 6.18. *Let $\mathcal{Q} \in \mathbf{CQ}^{\neq}[sm]$ be a strongly minimal conjunctive query with inequalities and let $\boldsymbol{P}$ be a distribution policy over an ordered network $\mathcal{N}$. $PC_{ord,set}(\boldsymbol{P}, \mathcal{Q})$ holds if and only if $PC_{ord,bag}(\boldsymbol{P}, \mathcal{Q})$ holds.*

Proof. (if) This direction is a trivial consequence of Proposition 6.16.

(only if) Assume $PC_{\text{ord,set}}(\boldsymbol{P}, \mathcal{Q})$ holds. In other words, for every minimal valuation V for $\mathcal{Q}$, there is a node $\kappa \in \mathcal{N}$ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. Since $\mathcal{Q}$ is strongly minimal, every valuation V for $\mathcal{Q}$ is minimal. Consequently, for every valuation V for $\mathcal{Q}$, there is a node $\kappa \in \mathcal{N}$ having $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. We conclude that $PC_{\text{ord,bag}}(\boldsymbol{P}, \mathcal{Q})$ holds as well. □

Notice that for a policy $\boldsymbol{P}$ and a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}$, $PC_{\text{ord,set}}(\boldsymbol{P}, \mathcal{Q})$ might coincide with $PC_{\text{ord,bag}}(\boldsymbol{P}, \mathcal{Q})$, even if $\mathcal{Q}$ is not strongly minimal. This is for example the case if $\boldsymbol{P}$ maps all the facts to the same node.

6.8.2 Transferability

When comparing the characterizations for transferability over ordered networks under set and bag semantics, described in respectively Proposition 6.5 and Proposition 6.14, it can easily be seen that the only significant difference between them is the notion of minimal valuations under set semantics. As one might expect, $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ doesn't necessarily imply $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$.

Example 6.19. As an example showing that $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ not necessarily implies $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$, consider the conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$,

$$\mathcal{Q} : H(x) \leftarrow R(x),$$
$$\mathcal{Q}' : H(x) \leftarrow R(x), R(y).$$

Obviously, a minimal valuation for $\mathcal{Q}'$ always maps the variables x and y onto the same value. Therefore, every minimal valuation V' for $\mathcal{Q}'$ requires

exactly one fact. It can easily be seen that there is a minimal valuation V for $\mathcal{Q}$ requiring this fact as well. Therefore, $\mathcal{Q}$ covers $\mathcal{Q}'$, so $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ holds.

Next, consider the valuation $V' = \{x \mapsto a, y \mapsto b\}$ for $\mathcal{Q}'$, with $a \neq b$. Since $V'(body_{\mathcal{Q}'})$ requires two different facts, there clearly is no valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$ as $body_{\mathcal{Q}}$ has only one atom. We conclude that $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ cannot hold. ■

In Section 5.2 we explained why $PC_{\text{bag}}(\boldsymbol{P}, \mathcal{Q})$ implying $PC_{\text{set}}(\boldsymbol{P}, \mathcal{Q})$ isn't sufficient to conclude that $PCT_{\text{bag}}(\mathcal{Q}, \mathcal{Q}')$ implies $PCT_{\text{set}}(\mathcal{Q}, \mathcal{Q}')$. This reasoning is applicable over ordered networks as well. In fact, if the condition for transferability over ordered networks under bag semantics is satisfied, it doesn't immediately follow that the condition for transferability over ordered networks under set semantics is satisfied. Indeed, if the former condition is satisfied, we know that for every valuation V' for $\mathcal{Q}'$ there has to be a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$. Consequently, for every minimal valuation V'_{min} for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ such that $V'_{\text{min}}(body_{\mathcal{Q}'}) \subseteq V(body_{\mathcal{Q}})$. Unfortunately, this doesn't necessarily imply that for every minimal valuation V'_{min} for $\mathcal{Q}'$ there is a minimal valuation V_{min} for $\mathcal{Q}$ with $V'_{\text{min}}(body_{\mathcal{Q}'}) \subseteq V_{\text{min}}(body_{\mathcal{Q}})$.

Example 6.20. As an example showing that $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ not necessarily implies $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$, consider the following conjunctive queries $\mathcal{Q}$ and $\mathcal{Q}'$,

$$\mathcal{Q} : H(x) \leftarrow R(x), R(y),$$
$$\mathcal{Q}' : H(x, y) \leftarrow R(x), R(y).$$

Since $body_{\mathcal{Q}} = body_{\mathcal{Q}'}$, it immediately follows that for every valuation V' for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{Q})$. We conclude that $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ holds.

Now consider the valuation $V' = \{x \mapsto a, y \mapsto b\}$ for $\mathcal{Q}'$. Since V' is a minimal valuation for $\mathcal{Q}'$, there has to be a minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{Q})$ for $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ to hold. In this case, the only possible valuations for $\mathcal{Q}$ are $V_1 = \{x \mapsto a, y \mapsto b\}$ and $V_2 = \{x \mapsto b, y \mapsto a\}$. Notice that these valuations V_1 and V_2 aren't minimal. Indeed, by assigning the same value as x to y, we get in both cases a valuation for $\mathcal{Q}$ deriving the same fact while requiring only one fact. We conclude that there is no minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_{Q})$. Consequently, $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ doesn't hold. ■

Notice that the only difference between bag and set semantics are minimal valuations. Therefore, it is possible to let transferability over ordered networks under set and bag semantics coincide by restricting ourselves to strongly minimal conjunctive queries.

Proposition 6.21. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be strongly minimal queries in $\mathbf{CQ}^{\neq}[sm]$. $PCT_{ord,set}(\mathcal{Q}, \mathcal{Q}')$ holds if and only if $PCT_{ord,bag}(\mathcal{Q}, \mathcal{Q}')$ holds.*

Proof. (if) Assume $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$. By assumption, for every valuation V' for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. Since $\mathcal{Q}$ and $\mathcal{Q}'$ are strongly minimal, it follows that for every minimal valuation V' for $\mathcal{Q}'$ there is a minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. Thus, $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ holds.

(only if) Assume $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$. By assumption, for every minimal valuation V' for $\mathcal{Q}'$ there is a minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. Since $\mathcal{Q}$ and $\mathcal{Q}'$ are strongly minimal, every valuation is minimal. Therefore, for every valuation V' for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. We conclude that $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ holds. □

Notice that it is not required for $\mathcal{Q}$ and $\mathcal{Q}'$ to be strongly minimal to let $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ and $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ be the same. As a trivial counterexample, consider the case where $\mathcal{Q}$ and $\mathcal{Q}'$ are arbitrary queries in $\mathbf{CQ}^{\neq}$ with $\mathcal{Q} = \mathcal{Q}'$. In this case, parallel-correctness clearly transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ (and vice versa), both under set and bag semantics.

We conclude our observation by considering the case where only $\mathcal{Q}$ is a strongly minimal query. In this case, there is still an implication from $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$ to $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$.

Proposition 6.22. *Let $\mathcal{Q}$ be a strongly minimal query in $\mathbf{CQ}^{\neq}[sm]$ and let $\mathcal{Q}'$ be a query in $\mathbf{CQ}^{\neq}$. $PCT_{ord,set}(\mathcal{Q}, \mathcal{Q}')$ holds if $PCT_{ord,bag}(\mathcal{Q}, \mathcal{Q}')$ holds.*

Proof. Assume $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$. By assumption, for every valuation V' for $\mathcal{Q}'$ there is a valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. Since $\mathcal{Q}$ is strongly minimal, it follows that for every valuation V' for $\mathcal{Q}'$ there is a minimal valuation V for $\mathcal{Q}$ with $V'(body_{\mathcal{Q}'}) \subseteq V(body_Q)$. Clearly, this holds for every minimal valuation V' for $\mathcal{Q}'$ as well, so $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ holds. □

Observe that while only $\mathcal{Q}$ is strongly minimal, $PCT_{\text{ord,set}}(\mathcal{Q}, \mathcal{Q}')$ not always follows from $PCT_{\text{ord,bag}}(\mathcal{Q}, \mathcal{Q}')$, as illustrated by Example 6.19.

Chapter 7

Hypercube distributions

A Hypercube distribution for a given conjunctive query $\mathcal{Q}$ partitions the data across the different nodes in an instance independent way based on the structure of $\mathcal{Q}$. This technique can be traced back to Ganguly, Silberschatz and Tsur [11] and is studied in the context of map-reduce by Afrati and Ullman [3]. Beame, Koutris and Suciu [7] referred to this technique as the Hypercube algorithm and used it to provide an upper bound for the amount of communication needed to compute a full conjunctive query without self-joins in one communication round.

In this chapter, we first define Hypercube distributions for a conjunctive query $\mathcal{Q}$. Next, we briefly summarize the results obtained by Ameloot et al. [4] related to Hypercube distributions under set semantics. We end this chapter with a study on the parallel-correctness of Hypercube distributions under bag semantics.

7.1 Definition

Let $\mathcal{Q}$ be a conjunctive query with $body_{\mathcal{Q}} = \{R_1(\mathbf{x}_1), \ldots, R_m(\mathbf{x}_m)\}$ and $vars(\mathcal{Q}) = \{x_1, \ldots, x_k\}$. Let $p_1, \ldots, p_k$ be strictly positive natural numbers and let $H = (h_1, \ldots, h_k)$ be a collection of hash functions with each hash function h_i mapping values from **dom** to values in $\{1, \ldots, p_i\}$. Assume a network $\mathcal{N}$ containing $p = p_1 \times \cdots \times p_k$ nodes. Each node in $\mathcal{N}$ is assigned a unique address in $\{1, \ldots, p_1\} \times \cdots \times \{1, \ldots, p_k\}$. Intuitively, these addresses organize the nodes in $\mathcal{N}$ in a hypercube of k dimensions with each dimension i having a size of p_i.

This set H of hash functions determines a hypercube distribution $\boldsymbol{P}_H$ for $\mathcal{Q}$ over $\mathcal{N}$ as follows. For each valuation V for $\mathcal{Q}$ and for each atom $A = R_i(\mathbf{x}_i)$ in $body_{\mathcal{Q}}$, the fact $f = V(A)$ is assigned to all the nodes in $\mathcal{N}$ having an address of the form $(b_1, \ldots, b_k)$ with $b_j = h_j(V(x_j))$ for all variables $x_j \in \mathbf{x}_i$. In other words, a fact $f = R(a_1, \ldots, a_l)$ is assigned to a node κ with address $(b_1, \ldots, b_k)$ if and only if f is mappable onto an atom A

in $body_{\mathcal{Q}}$ over the same relation R such that $b_i = h_i(a_r)$ if variable x_i appears in atom A on position r and $b_i \in img(h_i)$ otherwise for every $i \in \{1, \ldots, k\}$.

Example 7.1. Consider the following conjunctive query $\mathcal{Q}$,

$$H(x_1, x_3) \leftarrow R(x_1, x_2), R(x_2, x_3),$$

and let $p_1 = p_2 = 2$ and $p_3 = 1$. Assume for convenience that we are working under a binary universe $U = (a, b)$. Let $H = (h_1, h_2, h_3)$ be a collection of hash functions with $h_3(d) = 1$ for each value $d \in \mathbf{dom}$, $h_1(a) = h_2(a) = 1$ and $h_1(b) = h_2(b) = 2$.

Let $\mathcal{N}$ be a network over four nodes $\kappa_{1,1,1}$, $\kappa_{1,2,1}$, $\kappa_{2,1,1}$ and $\kappa_{2,2,1}$ having respectively the addresses $(1, 1, 1)$, $(1, 2, 1)$, $(2, 1, 1)$ and $(2, 2, 1)$. The distribution policy $\boldsymbol{P}_H$ distributes the facts over relation R as follows over the network $\mathcal{N}$:

- The fact R(a,a) is assigned to $\kappa_{1,1,1}$ because of the atom $R(x_1, x_2)$ and to both $\kappa_{1,1,1}$ and $\kappa_{2,1,1}$ because of atom $R(x_2, x_3)$.
- The fact R(a,b) is assigned to $\kappa_{1,2,1}$ because of the atom $R(x_1, x_2)$ and to both $\kappa_{1,1,1}$ and $\kappa_{2,1,1}$ because of atom $R(x_2, x_3)$.
- The fact R(b,a) is assigned to $\kappa_{2,1,1}$ because of the atom $R(x_1, x_2)$ and to both $\kappa_{1,2,1}$ and $\kappa_{2,2,1}$ because of atom $R(x_2, x_3)$.
- The fact R(b,b) is assigned to $\kappa_{2,2,1}$ because of the atom $R(x_1, x_2)$ and to both $\kappa_{1,2,1}$ and $\kappa_{2,2,1}$ because of atom $R(x_2, x_3)$.

Notice that this distribution policy $\boldsymbol{P}_H$ is not the only possible Hypercube distribution for $\mathcal{Q}$. It can easily be seen that a different collection of hash functions $H' = (h'_1, h'_2, h'_3)$ might result in a different Hypercube distribution $\boldsymbol{P}_{H'}$ for $\mathcal{Q}$. ■

7.2 Hypercube under set semantics

Let $\boldsymbol{P}_H$ be a hypercube distribution for a conjunctive query $\mathcal{Q}$. Ameloot et al. [4] proved that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}_H$. This Hypercube distribution $\boldsymbol{P}_H$ furthermore satisfies Condition 3.4. Indeed, a valuation V for $\mathcal{Q}$ is by definition of Hypercube distributions satisfiable on the node $\kappa \in \mathcal{N}$ with address $(h_1(V(x_1)), \ldots, h_k(V(x_k)))$.

Proposition 7.2 ([4])**.** *Let $\boldsymbol{P}_H$ be a hypercube distribution for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}$. Then $\boldsymbol{P}_H$ strongly saturates $\mathcal{Q}$.*

7.3 Hypercube under bag semantics

In contrast to set semantics, a conjunctive query $\mathcal{Q}$ is unfortunately not necessarily parallel-correct under a Hypercube distribution $\boldsymbol{P}_H$ for $\mathcal{Q}$ while considering bag semantics. We already mentioned that a Hypercube distribution $\boldsymbol{P}_H$ for a conjunctive query $\mathcal{Q}$ always satisfies Condition 3.4. This condition is however not sufficient for parallel-correctness under bag semantics, as Proposition 4.6 requires that there is exactly one node in the network responsible for each valuation V for $\mathcal{Q}$. However, the definition of a Hypercube distribution allows a valuation to be satisfiable on multiple nodes in the network.

Example 7.3. As a counterexample showing that under bag semantics a conjunctive query $\mathcal{Q}$ is not necessarily parallel-correct under every Hypercube distribution $\boldsymbol{P}_H$ for $\mathcal{Q}$, reconsider from Example 7.1 the conjunctive query $\mathcal{Q}$,

$$H(x_1, x_3) \leftarrow R(x_1, x_2), R(x_2, x_3),$$

and Hypercube distribution policy $\boldsymbol{P}_H$ based on the collection of hash functions $H = (h_1, h_2, h_3)$.

Let $V = \{x_1 \mapsto a, x_2 \mapsto a, x_3 \mapsto a\}$ be a valuation for $\mathcal{Q}$ over the binary universe $U = \{a, b\}$. As explained in Example 7.1, the fact $R(a, a)$ is mapped onto both $\kappa_{1,1,1}$ and $\kappa_{2,1,1}$. But $V(body_{\mathcal{Q}}) = \{R(a, a)\}$, so it immediately follows that $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}_H}(\kappa_{1,1,1})$ and $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}_H}(\kappa_{2,1,1})$. According to Proposition 4.6, $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}_H$ under bag semantics. ■

We next consider the application of Hypercube distributions under the modified distributed evaluation model, as described in Chapter 6. To this end, assume there is an arbitrary total order on the nodes in the network. We could, for example, sort the nodes in the network in a lexicographical order based on the address of each node.

Recall from Proposition 6.10 that a conjunctive query $\mathcal{Q}$ is parallel-correct under a distribution policy $\boldsymbol{P}$ over an ordered network $\mathcal{N}$ if and only if for every valuation V for $\mathcal{Q}$ there is at least one node κ in $\mathcal{N}$ with $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$. It can easily be seen that a Hypercube distribution for $\mathcal{Q}$ satisfies this condition.

Proposition 7.4. *Let $\boldsymbol{P}_H$ be a Hypercube distribution for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}$ over an ordered network $\mathcal{N}$. Then $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}_H$ over the ordered network $\mathcal{N}$ under bag semantics.*

Proof. Assume a Hypercube distribution $\boldsymbol{P}_H$ for a conjunctive query $\mathcal{Q} \in \mathbf{CQ}$ over an ordered network $\mathcal{N}$. According to Proposition 7.2, $\boldsymbol{P}_H$ strongly saturates $\mathcal{Q}$. In other words, Condition 3.4 is satisfied for $\mathcal{Q}$ and $\boldsymbol{P}_H$. It

clearly follows that Condition 6.8 is satisfied as well. According to Proposition 6.10, $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}_H$ over the ordered network $\mathcal{N}$ under bag semantics.

Alternatively, we could prove directly that Condition 6.8 is satisfied to conclude that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}_H$ over the ordered network $\mathcal{N}$ under bag semantics. To this end, assume an arbitrary valuation V for $\mathcal{Q}$. Let $vars(\mathcal{Q}) = \{x_1, \ldots, x_k\}$ and let $H = (h_1, \ldots, h_k)$. Now consider the node $\kappa \in \mathcal{N}$ with address $(h_1(V(x_1)), \ldots, h_k(V(x_k)))$. By definition of the Hypercube distribution $\boldsymbol{P}_H$, all the facts in $V(body_{\mathcal{Q}})$ are assigned to this node κ. In other words, $V(body_{\mathcal{Q}}) \subseteq rfacts_{\boldsymbol{P}_H}(\kappa)$. This reasoning is clearly valid for each valuation V for $\mathcal{Q}$, directly implying that Condition 6.8 is indeed satisfied for $\mathcal{Q}$ and $\boldsymbol{P}_H$. □

In Chapter 6, we showed that the characterization for parallel-correctness under set semantics doesn't change while considering ordered networks instead of regular networks. This implies in particular that under set semantics a Hypercube distribution $\boldsymbol{P}_H$ for a conjunctive query $\mathcal{Q}$ strongly saturates $\mathcal{Q}$ under the modified model as well. In other words, $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}_H$ over an ordered network under set semantics. We therefore conclude that parallel-correctness under Hypercube distributions over ordered networks coincides for set and bag semantics.

Chapter 8

Unions of conjunctive queries

In this chapter we study the extension of our results related to parallel-correctness and transferability under bag semantics toward unions of conjunctive queries with inequalities.

8.1 Parallel-correctness

Recall from Chapter 4 that Condition 4.3 is both necessary and sufficient for parallel-correctness in the context of conjunctive queries with inequalities under bag semantics. We modify Condition 4.3 to include unions of conjunctive queries with inequalities.

Condition 8.1. *Let $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ be a union of conjunctive queries with inequalities and $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ a distribution policy over a network $\mathcal{N}$. For every valuation V_i for $\mathcal{Q}$ over U, witnessed by some $\mathcal{Q}_i \in \mathcal{Q}$, there is exactly one node $\kappa \in \mathcal{N}$ such that $V_i(body_{\mathcal{Q}_i}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Before proving that Condition 8.1 is a necessary and sufficient condition for parallel-correctness for unions of conjunctive queries with inequalities under bag semantics, we first mention that Lemma 4.4 and Lemma 4.5 are applicable for unions of conjunctive queries as well.

Lemma 8.2. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is not parallel-correct under $\boldsymbol{P}$ if there exists a valuation V_i for $\mathcal{Q}$, witnessed by some query $\mathcal{Q}_i \in \mathcal{Q}$, and more than one node $\kappa \in \mathcal{N}$ with $V_i(body_{\mathcal{Q}_i}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proof sketch. The proof is analogous to the proof of Lemma 4.4. Assume towards a contradiction that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. Let κ_1 and κ_2 be two different nodes in $\mathcal{N}$ with $V_i(body_{\mathcal{Q}_i}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_1)$ and $V_i(body_{\mathcal{Q}_i}) \subseteq rfacts_{\boldsymbol{P}}(\kappa_2)$. Consider an instance I with $facts(I) = V_i(body_{\mathcal{Q}_i})$. It can easily be seen that all the facts in I are mapped onto both κ_1 and κ_2, meaning that all the results in $\mathcal{Q}(I)$ are derived on both nodes. In particular, the fact

$f = V(head_{\mathcal{Q}_i})$ will be derived on both nodes, implying that the resulting multiplicity of f in the distributed evaluation will be too high, unless there is some kind of compensation. The only possible way to compensate is the existence of a valuation V_j, witnessed by some $\mathcal{Q}_j \in \mathcal{Q}$, contributing to the multiplicity of f in $\mathcal{Q}(I)$, but not to the multiplicity of f in $[\mathcal{Q}, \boldsymbol{P}](I)$. This is however impossible, as the required facts for V_j would clearly be available on both κ_1 and κ_2. □

Lemma 8.3. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is not parallel-correct under $\boldsymbol{P}$ if there exists a valuation V_i for $\mathcal{Q}$, witnessed by some query $\mathcal{Q}_i \in \mathcal{Q}$, with no node $\kappa \in \mathcal{N}$ with $V_i(body_{\mathcal{Q}_i}) \subseteq rfacts_{\boldsymbol{P}}(\kappa)$.*

Proof sketch. The proof is analogous to the proof of Lemma 4.5. Assume towards a contradiction that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. Consider an arbitrary instance I containing the required facts for V_i. Since no node in the network contains all the required facts for V_i, it can easily be seen that the multiplicity of the fact f derived by V_i will be too low in the distributed result, unless there is some compensation in the form of a valuation V_j for $\mathcal{Q}$, witnessed by a query $\mathcal{Q}_j \in \mathcal{Q}$, deriving f on multiple nodes. The latter contradicts with Lemma 8.2, implying that $\mathcal{Q}$ cannot be parallel-correct under $\boldsymbol{P}$. □

We use both lemmas to prove that Condition 8.1 is a necessary and sufficient condition for parallel-correctness for queries in $\mathbf{UCQ}^{\neq}$ under bag semantics.

Proposition 8.4. *Let $\boldsymbol{P} = (U, rfacts_{\boldsymbol{P}})$ be a distribution policy over a network $\mathcal{N}$. A query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is parallel-correct under $\boldsymbol{P}$ if and only if Condition 8.1 is satisfied.*

Proof sketch. (if) Assume Condition 8.1 holds. Clearly, every valuation for $\mathcal{Q}$ is satisfiable on exactly one node, implying that $\mathcal{Q}(I)$ equals $[\mathcal{Q}, \boldsymbol{P}](I)$ for every instance I.

(only if) The proof is by contraposition. Assume Condition 8.1 doesn't hold. It immediately follows from Lemma 8.2 and Lemma 8.3 that $\mathcal{Q}$ is not parallel-correct under $\boldsymbol{P}$. □

Since the characterization for parallel-correctness didn't change while considering unions of conjunctive queries instead of conjunctive queries, the complexity of deciding parallel-correctness under bag semantics remains the same as well.

Proposition 8.5. *The problem $\mathbf{PC}(\mathbf{UCQ}^{\neq}, \mathcal{P})$ is in Π_2^p for every distribution policy class $\mathcal{P} \in \{\mathcal{P}_{fin}\} \cup \mathfrak{P}_{nondet}$.*

Proof sketch. The construction of a Π_2^p-algorithm is analogous to the algorithm described in the proof of Proposition 4.7. The only difference is that valuations for $\mathcal{Q}$ are now witnessed by some query $\mathcal{Q}_i \in \mathcal{Q}$. □

Proposition 8.6. *The problem* $\mathbf{PC}(\mathbf{UCQ}^{\neq}, \mathcal{P})$ *is in* coNP *for every distribution policy class* $\mathcal{P} \in \mathfrak{P}_{det} \cup \{\mathcal{P}_{fin}\}$.

Proof sketch. Analogously to the proof of Proposition 4.8, we show that $\overline{\mathbf{PC}(\mathbf{UCQ}^{\neq}, \mathcal{P}^{\mathbf{k}}_{\mathbf{det}})}$ is in NP. The construction of a nondeterministic polynomial algorithm is as follows: guess a valuation V and check if the number of nodes containing all the required facts for V is different from 1. The proof of Proposition 4.8 already explained that this algorithm is indeed executable in polynomial time. □

Notice that the lower bound provided in Proposition 4.9 is a lower bound for parallel-correctness under unions of conjunctive queries with inequalities as well, since $\mathbf{CQ} \subseteq \mathbf{UCQ}^{\neq}$.

8.2 Parallel-correctness transfer

The definition of *impFacts* extends to unions of conjunctive queries with inequalities in a natural way:

Definition 8.7. Let V_i be a valuation for a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ over a universe U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$. A fact f over U is in $\mathit{impFacts}(V_i, \mathcal{Q})$ if and only if for every distribution policy $\boldsymbol{P} = (U, \mathit{rfacts}_{\boldsymbol{P}})$ over a network $\mathcal{N}$ under which $\mathcal{Q}$ is parallel-correct and for every node $\kappa \in \mathcal{N}$, if $V_i(\mathit{body}_{\mathcal{Q}_i}) \subseteq \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$, then $f \in \mathit{rfacts}_{\boldsymbol{P}}(\kappa)$.

The properties and inference rules for *impFacts* described in Section 4.3.2 are furthermore applicable to unions of conjunctive queries with inequalities as well. The only difference for queries in $\mathbf{UCQ}^{\neq}$ instead of $\mathbf{CQ}^{\neq}$ is that a valuation V for a query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is now witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$. This difference however doesn't influence the different proof ideas in Section 4.3.2.

Next, we modify Condition 4.26 to include unions of conjunctive queries with inequalities:

Condition 8.8. *Let* $\mathcal{Q}$ *and* $\mathcal{Q}'$ *be queries in* $\mathbf{UCQ}^{\neq}$. *For each valuation* V' *for* $\mathcal{Q}'$ *over a universe* U, *witnessed by a query* $\mathcal{Q}'_j \in \mathcal{Q}'$, *there exists a valuation* V *for* $\mathcal{Q}$ *over* U, *witnessed by a query* $\mathcal{Q}_i \in \mathcal{Q}$, *such that* $V(\mathit{body}_{\mathcal{Q}_i}) \subseteq V'(\mathit{body}_{\mathcal{Q}'_j}) \subseteq \mathit{impFacts}(V, \mathcal{Q}_i)$.

Condition 8.8 is a necessary and sufficient condition for transferability. Before proving this proposition, we first reformulate Lemma 4.28, as it is applicable to queries in $\mathbf{UCQ}^{\neq}$ as well.

Lemma 8.9. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in* $\mathbf{UCQ}^{\neq}$*. Condition 8.8 isn't satisfied, witnessed by some valuation V' for a query $\mathcal{Q}'_j \in \mathcal{Q}'$ over a universe U, if and only if there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j})$ or there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$.*

Proof sketch. (if) Let V' be a valuation as described in Lemma 4.28. For both cases, it immediately follows that there cannot exist a valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ With $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$. Therefore, Condition 4.26 isn't satisfied.

(only if) Assume Condition 8.8 isn't satisfied, witnessed by some valuation V' for a query $\mathcal{Q}'_j \in \mathcal{Q}'$ over a universe U. By assumption, every valuation V for a query $\mathcal{Q}_i \in \mathcal{Q}$ over U satisfies one of the following three conditions:

1. $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j}) \not\subseteq impFacts(V, \mathcal{Q}_i)$
2. $V(body_{\mathcal{Q}_i}) \not\subseteq V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$
3. $V(body_{\mathcal{Q}_i}) \not\subseteq V'(body_{\mathcal{Q}'_j}) \not\subseteq impFacts(V, \mathcal{Q}_i)$

As explained in the proof of Lemma 4.28, condition 1 and condition 3 cannot occur together, thereby proving that there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j})$ or there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$. □

Proposition 8.10. *Let $\mathcal{Q}$ and $\mathcal{Q}'$ be queries in* $\mathbf{UCQ}^{\neq}$*. Parallel-correctness transfers from $\mathcal{Q}$ to $\mathcal{Q}'$ if and only if Condition 8.8 is satisfied.*

Proof sketch. (if) Assume Condition 8.8 holds. Let $\boldsymbol{P}$ be an arbitrary distribution policy under which $\mathcal{Q}$ is parallel-correct. We prove that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$ as well. To this end, let V' be an arbitrary valuation for a query $\mathcal{Q}_j \in \mathcal{Q}'$. By assumption, there is a valuation V for $\mathcal{Q}$, witnessed by some query $\mathcal{Q}_i \in \mathcal{Q}$, with $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$. As explained in the proof of Proposition 4.29, it follows that there is exactly one node in the network containing all the required facts for V'. We conclude that $\mathcal{Q}'$ is parallel-correct under $\boldsymbol{P}$.

(only if) The proof is by contraposition. Assume Condition 8.8 doesn't hold. We show that parallel-correctness doesn't transfer from $\mathcal{Q}$ to $\mathcal{Q}'$ by constructing distribution policies under which $\mathcal{Q}$ is parallel-correct, but $\mathcal{Q}'$ is not. According to Lemma 8.9, there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j})$ or there is no

valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$.

First, consider the case where there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V(body_{\mathcal{Q}_i}) \subseteq V'(body_{\mathcal{Q}'_j})$. In this case, we construct $\boldsymbol{P}$ over two nodes κ_1 and κ_2 by assigning each fact to κ_1 and the required facts for V' to κ_2 as well. As explained in the proof of Proposition 4.29, $\mathcal{Q}$ is parallel-correct under this distribution policy, but $\mathcal{Q}'$ clearly isn't.

Next, consider the case where there is no valuation V for $\mathcal{Q}$ over U, witnessed by a query $\mathcal{Q}_i \in \mathcal{Q}$ with $V'(body_{\mathcal{Q}'_j}) \subseteq impFacts(V, \mathcal{Q}_i)$. It is furthermore safe to assume the existence of a valuation V_1 for a query $\mathcal{Q}_k \in \mathcal{Q}$ with $V_1(body_{\mathcal{Q}_k}) \subseteq V'(body_{\mathcal{Q}'_j})$, because $\mathcal{Q}$ and $\mathcal{Q}'$ are already covered by the first case if this valuation V_1 doesn't exist. By assumption, $V'(body_{\mathcal{Q}'_j}) \not\subseteq impFacts(V_1, \mathcal{Q})$. This implies the existence of a distribution policy $\boldsymbol{P}$ under which $\mathcal{Q}$ is parallel-correct and where the required facts for V' do not meet on the same node as those for V_1. We show in the proof of Proposition 4.29 that $\mathcal{Q}'$ cannot be parallel-correct under this distribution policy $\boldsymbol{P}$. □

The complexity of deciding transferability under bag semantics doesn't alter when considering unions of conjunctive queries instead of conjunctive queries.

Proposition 8.11. *The problem* **PC-Trans**($\mathbf{UCQ}^{\neq}, \mathbf{UCQ}^{\neq}$) *is in* EXPTIME.

Proof sketch. The construction of an EXPTIME-algorithm deciding transferability for queries in $\mathbf{UCQ}^{\neq}$ is analogous to the constructed algorithm in the proof of Proposition 4.31. Instead of considering all possible valuations V and V' for respectively queries $\mathcal{Q}, \mathcal{Q}' \in \mathbf{CQ}^{\neq}$, we now need to consider all possible valuations V_i and V'_j for each subquery $\mathcal{Q}_i \in \mathcal{Q}$ and $\mathcal{Q}'_j \in \mathcal{Q}'$, with $\mathcal{Q}, \mathcal{Q}' \in \mathbf{UCQ}^{\neq}$. □

Next, consider the class of unions of conjunctive queries with inequalities without self-joins, denoted $\mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$. A query $\mathcal{Q} \in \mathbf{UCQ}^{\neq}$ is in $\mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$ if for every subquery $\mathcal{Q}_i \in \mathcal{Q}$ it holds that $\mathcal{Q}_i \in \mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$. Proposition 4.32 states that for a query $\mathcal{Q} \in \mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$, the set $impFacts(V, \mathcal{Q})$ equals $V(body_{\mathcal{Q}})$ for every valuation V for $\mathcal{Q}$. Unfortunately, this result doesn't necessarily hold for queries in $\mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$.

Example 8.12. As a counterexample showing that $impFacts(V, \mathcal{Q})$ not necessarily equals $V(body_{\mathcal{Q}})$ for every valuation V for a query $Q \in \mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$, consider the query $\mathcal{Q} = \mathcal{Q}_1 \cup \mathcal{Q}_2$, with

$$\mathcal{Q}_1 : H(x) \leftarrow R(x),$$
$$\mathcal{Q}_2 : H(y) \leftarrow R(y), S(y).$$

Now consider the valuation V_1 for $\mathcal{Q}_1$, mapping x to 1, and the valuation V_2 for $\mathcal{Q}_2$, mapping y to 1. Since $V_1(body_{\mathcal{Q}_1}) = \{R(1)\}$ and $V_2(body_{\mathcal{Q}_2}) = \{R(1), S(1)\}$, it follows that $V_1(body_{\mathcal{Q}_1}) \subseteq V_2(body_{\mathcal{Q}_2})$. Therefore, the required facts for V_1 and V_2 should always meet at the same node, since V_1 would otherwise be satisfiable on both nodes. We conclude that $S(1) \in \mathit{impFacts}(V_1, \mathcal{Q})$, implying that $\mathit{impFacts}(V_1, \mathcal{Q}) \neq V_1(body_{\mathcal{Q}_1})$. ■

We can achieve the desired result by adding an extra constraint on the queries in $\mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$. The construction in the previous example illustrated the fact that the required facts for a valuations for one subquery can be a subset of the required facts for a valuation for another subquery. We prevent this possibility by furthermore requiring that each relation name is used at most once in the query as a whole. Notice that for conjunctive queries without self-joins this requirement is always fulfilled. We use $\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}}$ to denote this subset of $\mathbf{UCQ}^{\neq}$ satisfying the condition that every relation appears at most once in the whole query[1]. It can easily be seen that $\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}} \subseteq \mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$.

Proposition 8.13. *Let $\mathcal{Q}$ be a query in $\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}}$. For every valuation V for a query $\mathcal{Q}_i \in \mathcal{Q}$ over a universe U, $\mathit{impFacts}(V, \mathcal{Q}) = V(body_{\mathcal{Q}_i})$.*

Proof. Let $\mathcal{Q}$ be a query in $\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}}$. Consider an arbitrary valuation V for a subquery $\mathcal{Q}_i \in \mathcal{Q}$ over a universe U. We prove that for every fact f over U, it holds that $f \in \mathit{impFacts}(V, \mathcal{Q})$ if and only if $f \in V(body_{\mathcal{Q}_i})$.

(if) Assume $f \in V(body_{\mathcal{Q}_i})$. For every distribution policy $\boldsymbol{P}$ over U under which $\mathcal{Q}$ is parallel-correct, f trivially appears on the same node as the facts in $V(body_{\mathcal{Q}_i})$. Consequently, $f \in \mathit{impFacts}(V, \mathcal{Q})$.

(only if) The proof is by contraposition. Assume $f \notin V(body_{\mathcal{Q}_i})$. We show that $f \notin \mathit{impFacts}(V, \mathcal{Q})$ by constructing a distribution policy $\boldsymbol{P}$ over U such that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$ and f does not appear on the node responsible for V.

Since $\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}} \subseteq \mathbf{UCQ}^{\neq}_{\neg\mathbf{sj}}$, it follows that $\mathcal{Q}_i \in \mathbf{CQ}^{\neq}_{\neg\mathbf{sj}}$. Therefore, we can reuse the distribution policy constructed in the proof of Proposition 4.32. This proof describes a construction for a distribution policy $\boldsymbol{P}$ that can be used to isolate the facts in $V(body_{\mathcal{Q}_i})$ onto a separate node, while still making sure that $\mathcal{Q}_i$ is parallel-correct under $\boldsymbol{P}$. We extend this construction with an extra node κ_j to allow $\mathcal{Q}$ as a whole to be parallel-correct under $\boldsymbol{P}$ as well. We simply assign all the facts over a relation not in $\mathcal{Q}_i$ to this node κ_j. Since a subquery $\mathcal{Q}_j \in \mathcal{Q}$ different from $\mathcal{Q}_i$ cannot have a relation in common with $\mathcal{Q}_i$, it clearly follows that each valuation for $\mathcal{Q}_j$ over U is only satisfiable on κ_j. valuations for $\mathcal{Q}_i$ on the other hand can never be satisfied on this node. Therefore, it can easily be seen that $\mathcal{Q}$ is parallel-correct under $\boldsymbol{P}$. We conclude that $f \notin \mathit{impFacts}(V, \mathcal{Q})$. □

[1] In this notation, **dr** stands for duplicate relations.

Proposition 8.13 is useful to lower the complexity of transferability as follows:

Proposition 8.14. *The problem* **PC-Trans**$(\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}}, \mathbf{UCQ}^{\neq})$ *is in* Π^p_2.

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be the input queries for **PC-Trans**$(\mathbf{UCQ}^{\neq}_{\neg\mathbf{dr}}, \mathbf{UCQ}^{\neq})$. According to Proposition 8.13, it suffices to show that there is a Π^p_2-algorithm that checks if for each valuation V' for a query $\mathcal{Q}'_j \in \mathcal{Q}'$ over a universe U, there exists a valuation V for $\mathcal{Q}$ over U, witnessed by some query $\mathcal{Q}_i \in \mathcal{Q}$, such that $V(body_{\mathcal{Q}_i}) = V'(body_{\mathcal{Q}'_j})$. Since $V(body_{\mathcal{Q}_i}) = V'(body_{\mathcal{Q}'_j})$ can obviously be checked in polynomial time, the construction of the Π^p_2-algorithm is trivial. □

Chapter 9

Conclusion

In this thesis, we studied the possibility to extend parallel-correctness and transferability toward conjunctive queries under bag semantics. This extension is useful, as bag semantics is required to correctly perform certain aggregation functions like counting or averaging the results.

We first provided two conditions based on valuations that are both necessary and sufficient for parallel-correctness and transferability. Based on these characterizations, we provided upper bounds for the time complexity of deciding parallel-correctness and transferability for conjunctive queries with inequalities under bag semantics. The upper bound for deciding parallel-correctness is further improved by restricting the considered classes of distribution policies. For the latter case, we provided a matching lower bound as well. The upper bound for deciding transferability on the other hand is lowered by restricting the considered classes of conjunctive queries to queries without self-joins.

We then studied the relation of parallel-correctness and transferability between set and bag semantics. In particular, parallel-correctness under bag semantics always implies parallel-correctness under set semantics, whereas the converse is not necessarily true. We showed that parallel-correctness under set and bag semantics coincide if we limit ourselves to strongly minimal conjunctive queries and non-replicating distribution policies. Although parallel-correctness under bag semantics implies parallel-correctness under set semantics, transferability under bag semantics does not necessarily imply transferability under set semantics or vice versa.

The characterization for parallel-correctness under bag semantics appeared to be quite restrictive on possible distribution policies. Depending on the given conjunctive query, it might even be impossible to parallelize the evaluation of this query over more than one node in the network. We therefore introduced a modified distributed evaluation model based on ordered networks. The main advantage of this approach is that each valuation will derive a fact at most once, even if multiple nodes in the network

contain the required facts for this valuation. This modification doesn't alter the result under set semantics, meaning that the characterizations for parallel-correctness and transferability under set semantics still apply. The characterization for parallel-correctness under bag semantics on the other hand is simplified, as it is no longer required that each valuation for a given conjunctive query is satisfiable on exactly one node in the network. This simplification resulted in a different characterization for transferability under bag semantics as well and therefore allowed us to provide an improved upper bound on deciding transferability.

Next, we studied the application of Hypercube distributions for conjunctive queries under bag semantics. Unlike set semantics, queries are not necessarily parallel-correct under these Hypercube distributions while considering bag semantics. We showed on the other hand that conjunctive queries are indeed parallel-correct under their respective Hypercube distributions under bag semantics if we use the modified distributed evaluation model over ordered networks instead.

Lastly, we considered parallel-correctness and transferability under bag semantics for unions of conjunctive queries with inequalities. We showed that our obtained results can indeed be extended to this class of queries.

Bibliography

[1] Apache Hadoop. `http://hadoop.apache.org/`.

[2] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[3] F. N. Afrati and J. D. Ullman. Optimizing joins in a map-reduce environment. In *EDBT*, volume 426 of *ACM International Conference Proceeding Series*, pages 99–110. ACM, 2010.

[4] T. J. Ameloot, G. Geck, B. Ketsman, F. Neven, and T. Schwentick. Parallel-correctness and transferability for conjunctive queries. In *PODS*, pages 47–58. ACM, 2015. Full version to appear in Journal of the ACM.

[5] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia. Spark SQL: relational data processing in spark. In *SIGMOD Conference*, pages 1383–1394. ACM, 2015.

[6] P. Beame, P. Koutris, and D. Suciu. Communication steps for parallel query processing. In *PODS*, pages 273–284. ACM, 2013.

[7] P. Beame, P. Koutris, and D. Suciu. Skew in parallel query processing. In *PODS*, pages 212–223. ACM, 2014.

[8] S. Chaudhuri and M. Y. Vardi. Optimization of *Real* conjunctive queries. In *PODS*, pages 59–70. ACM Press, 1993.

[9] S. A. Cook. The complexity of theorem-proving procedures. In *STOC*, pages 151–158. ACM, 1971.

[10] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[11] S. Ganguly, A. Silberschatz, and S. Tsur. Parallel bottom-up processing of datalog queries. *J. Log. Program.*, 14(1&2):101–126, 1992.

[12] G. Geck, B. Ketsman, F. Neven, and T. Schwentick. Parallel-correctness and containment for conjunctive queries with union and negation. In *ICDT*, volume 48 of *LIPIcs*, pages 9:1–9:17. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2016.

[13] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Plenum Press, New York, 1972.

[14] P. Koutris and D. Suciu. Parallel evaluation of conjunctive queries. In *PODS*, pages 223–234. ACM, 2011.

[15] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *SIGMOD Conference*, pages 1099–1110. ACM, 2008.

[16] L. J. Stockmeyer. The polynomial-time hierarchy. *Theor. Comput. Sci.*, 3(1):1–22, 1976.

[17] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive - A warehousing solution over a map-reduce framework. *PVLDB*, 2(2):1626–1629, 2009.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Parallel-correctness and transferability for conjunctive queries under bag semantics

Richting: **master in de informatica-databases**
Jaar: **2017**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Vandevoort, Brecht

Datum: **6/06/2017**