The Uncertainty in Exceptional Model Mining

The Uncertainty in Exceptional Model Mining

Xin Du

The Uncertainty in Exceptional Model Mining by Xin Du

An electronic version of this dissertation is available at
http://research.tue.nl/.

ISBN: 978-90-386-5110-1

SIKS Dissertation Series No. 2020-25.
The research reported in this dissertation has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Exceptional Model Mining, Subgroup Discovery, Causal Inference, Model Evaluation, Probabilistic Graphic Model, Bayesian Modeling, Neural Networks.

Printed by: Ipskamp Printing

The Uncertainty in Exceptional Model Mining

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op donderdag 28 September 2020 om 11:00 uur door

Xin Du

geboren te Hebi, China

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. J.J. Lukkien	Technische Universiteit Eindhoven
1^{st} promotor:	prof.dr. M. Pechenizkiy	Technische Universiteit Eindhoven
copromotor(en):	dr. W. Duivesteijn	Technische Universiteit Eindhoven
leden:	prof.dr. D. Pedreschi	Università di Pisa
	prof.dr. A. Termier	Université de Rennes 1
	dr. A. Di Bucchianico	Technische Universiteit Eindhoven
	dr. M. Plantevit	Université Claude Bernard Lyon 1
adviseur(s):	dr. E. Galbrun	University of Eastern Finland

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

To my dear parents

Table of Contents

Summary

Exceptional Model Mining (EMM) is a local pattern mining approach that cares about how differently a model would perform in subpopulations, as compared to the same model but fitted on the whole data. Subgroup, model class, quality measure and search algorithm are the four essential components for EMM.

Assume a dataset consists of a set of descriptive variables and a set of target variables, a subgroup is a subset of data that are covered by a description defined in terms of descriptive variables. A model class is an arbitrary model defined on the target variables, e.g., regression model or classification model. A quality measure is a function that assigns a numeric value to a description, quantifying the difference of model's performance on the whole data and the subgroups supporting the description. A search algorithm can be guided by a quality measure to explore the space of descriptions (defined over descriptive variables) evaluated in terms of the performance of the chosen model , which allows us to find the top-Q (Q is an user-defined integer) exceptional subgroups.

In this dissertation, we study the problem of EMM with a focus on the uncertainty. We are particularly interested in studying the underlying mechanisms that determine the exceptionality of subgroups with observational datasets. By understanding such mechanisms, we are able to capture the uncertainty in EMM. In EMM, the fundamental assumption is that the exceptional performance of a model on a subgroup is governed by the dependency between target variables and by the dependency between descriptive variables and target variables. Description language, model class, quality measure are the three essential parts that determine the computation of exceptionality scores; search algorithm is the essential part that identifies the top exceptional subgroups guided by the computed exceptional scores. We develop several probabilistic models and employ statistical methods to quantify the exceptionality considering the uncertainty in dependency modeling with limited records. With several practical applications, we show how our methods can help users understand the different forms of exceptional behavior. Specifically, we propose to study this problem in four aspects:

- Uncertainty in multi-modal dependency modeling. When target variables consist of multi-modal interactions, such as sptial, temporal and word topics, it is difficult to capture the exceptionality of behavior in subgroups. Our main contribution is to encode the uncertainty in multi-

modal dependency by explicitly modeling the data generating process, and to provide a Bayesian inference method to estimate the latent factors in the generating process. In this process, we propose new a model class for nulti-modal interactions. By comparing the posterior distributions of the model parameters, we propose a quality measure to capture the exceptionality of multi-modal behavior in subgroups (Du et al., 2020b).

- Uncertainty in high-dimensional and heterogeneous dependency modeling. Our main contribution is proposing tools to model the complicated interactions between target variables, especially when the dimensionality of targets is very high. The aim is to help user understand how these complicated interactions could reflect the running mechanism of a model and provide explainable knowledge for people to improve real-world applications. We propose new model classes that explicitly represent the dependencies between variables and involved uncertainties. The practical studies are two-fold: on the one hand, our research is applied to discover subgroups of students, whose study behavior is exceptionally different from study behavior of those students in the whole data (Du et al., 2018). This provides insights to the practitioners in educational areas to make efficient policies improving the study grades of students. On the other hand, our research on fairness in network representation learning suggests that current network representation models like Node2vec (Grover and Leskovec, 2016), Deepwalk (Perozzi et al., 2014), would lead to biased performance on those disadvantaged subgroups (Du et al., 2020c). A further improvement is needed to ensure the learning of both a fair and structure-preserving network representation model.
- Uncertainty in individual causal dependency. Due to the observational equivalence in the historical datasets, our algorithm might report spurious exceptional subgroups. It is required to model the causal dependency and capture the uncertainty in causal relations. In order to solve the problem of predicting treatment outcome with observational data, we propose a neural network framework implementing the potential outcome framework (Rosenbaum and Rubin, 1983). Our work can be applied to evaluate whether a policy or clinic treatment is sufficiently effective with historical data. This can prevent the high costs of doing A/B tests or randomizing controlled trials (RCTs) (Du et al., 2019).
- Uncertainty in local causal dependency. Previous research either focuses on modeling the dependencies between target variables or assumes all description variables are correlated with all target variables. In this

study, we argue that properly capturing the uncertainty in the relation between attributes and targets can thoroughly boost the EMM performance. We call this relation Local Causal Dependency (LCD) and encode it with a causal graph language. We propose D-graph, a causal graph with extra nodes pointing to descriptive variables, which indicate the change of local mechanisms. We argue that the changing of local mechanisms in a causal graph lead to the changing of model's performance (Du et al., 2020a). This method can prevent the algorithm from searching the description space where the attributes do not influence the performance of the model, which could substantially boost the EMM process. At the same time, this method can improve the efficiency of retrieving non-redundant exceptional subgroups.

1 Introduction

"A tautology's truth is certain, a proposition's possible, a contradiction's impossible. Certain, possible, impossible: here we have the first indication of the scale that we need in the theory of probability."

Tractatus Logico-Philosophicus,
Ludwig Wittgenstein, 1922.

1.1 Background

In this chapter, we briefly review the Local Pattern Mining methods (Berthold et al., Hand, 2002) such as Subgroup Discovery (SD) (Atzmueller, 2015) and introduce the Exceptional Model Mining (EMM) (Duivesteijn et al., 2016, Leman et al., 2008) on which we focus throughout the dissertation. we demonstrate the main research questions, motivations and contributions by introducing the uncertainty in different forms of dependency modeling. Insights with several practical studies are provided, such as multi-modal dependency modeling and Local Causal Dependency (LCD) modeling.

1.1.1 Local Pattern Mining

The increasing amounts of data bring both new opportunities and challenges for data mining / machine learning research. From the aspect of opportunity, data mining / machine learning techniques could be broadly applied to real-world scenarios such as autonomous driving (Levinson et al., 2011), business intelligence (Negash and Gray, 2008), health care (Dua et al., 2014), finance (Kotsiantis et al., 2006), sports (Silver et al., 2017) and video games (Vinyals et al., 2019); From the aspect of challenge, traditional mathematical models are required to evolve from ideal assumptions to complex

real-world data with missing information, imbalanced distribution or noisy labels. However, it is nearly impossible to tackle all the challenges with one universal model (Wolpert and Macready, 1997). The possible solution is to evolve the developed dedicated methodology, tackling the challenges step by step with specific tasks. One of the most important tasks is to extract useful information from the large amount of data. By referring to useful information, here we mean useful for downstream tasks like classification and clustering, or for other tasks like multi-task learning and domain adaptation, or to better understand the data (Goodfellow et al., 2016). In particular, the useful information is demonstrated by patterns of interest with specific form of representations (Duivesteijn et al., 2016).

The specific form of representation we are going to introduce is called local pattern. Local Pattern Mining (LPM) (Hand, 2002, Morik et al., 2005) is a subfield of data mining, focused on discovering subsets of the dataset at hand which are interesting with regard to some quality measures. Typically, a restriction is imposed on what kind of interesting subsets we are looking for: only those subsets that can be formulated within a predefined *description language* are allowed. A common choice for this language is conjunctions of conditions on attributes of the dataset. Hence, if the records covered by a description is interesting for people, then results for LPM is shown in the form:

$$\text{Age} \geq 45 \wedge \text{Smoker} = \text{yes} \quad \rightsquigarrow \quad \text{interesting}$$

This ensures that the results we find with an LPM method are relatively easy to interpret for a domain expert: the subsets will be expressed in terms of quantities with which the expert is familiar. We call a subset that can be expressed in such a way a *subgroup*. Specifically, we assume a dataset Ω: a bag of N records $r \in \Omega$ of the form $r = (a_1, \ldots a_k, l_1, \ldots, l_m)$, where k and m are positive integers. We call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $l_1, \ldots, l_m$ the *target attributes* or *targets* of r. The descriptive attributes are taken from an unrestricted domain $\mathcal{A}$. Mathematically, we define descriptions as functions $D : \mathcal{A} \to \{0, 1\}$. A description D covers a record r^i if and only if $D(a_1^i, \cdots, a_k^i) = 1$.

Definition 1.1.1 A subgroup corresponding to a description D is the bag of records $S_D \subseteq \Omega$ that D covers, i.e.:

$$S_D = \left\{ r^i \in \Omega \middle| D(a_1^i, \ldots, a_k^i) = 1 \right\}$$

This merely formalizes the standard LPM conditions: we seek subgroups that are defined in terms of conditions on the descriptors, hence our results are

interpretable. Those conditions select a subset of the records of the dataset: those records that satisfy all conditions.

1.1.2 Subgroup Discovery

With regard to the different forms of interestingness measures, different LPM methods can give different results of subgroups. The most famous form of LPM is *Frequent Itemset Mining* (FIM) (Agrawal et al., 1996), where interestingness is measured by the exceptional frequency of occurrence: records that occur more frequently than a chosen threshold are considered interesting. Hence, FIM finds results of the form:

$$\text{Age} \geq 45 \wedge \text{Smoker} = \text{yes} \quad \rightsquigarrow \quad \text{(high frequency)}$$

If we are particularly interested in the patterns related to one target, then we need to reformulate the interestingness measure. The task of SD (Klösgen, 1996, Wrobel, 1997, Herrera et al., 2011) typically focuses on one binary attribute of the dataset as the *target*: subgroups are regarded as interesting if this one target has an unusual distribution, as compared to its distribution on the whole dataset. In our example, if the target column describes whether the person develops lung cancer or not, SD finds results of the form:

$$\text{Smoker} = \text{yes} \rightsquigarrow \text{lung cancer} = \text{yes}$$
$$\text{Age} \leq 25 \rightsquigarrow \text{lung cancer} = \text{no}$$

These subgroups make intuitive sense in terms of our knowledge of the domain. Smokers have a higher-than-usual incidence of lung cancer. People under the age of 25 often have low chance to develop lung cancer, so the incidence in this group will be lower. When the connection between subgroup and unusual target distribution is not immediately intuitively clear, the result of SD is a new hypothesis to be investigated by the domain experts.

1.1.3 Exceptional Model Mining

In the general concept of local pattern mining and subgroup discovery which cares about the particular distribution of a single target variable, we focus on two paradigms: summarization and distinctness detection. On the one hand, by defining a subgroup, we are able to represent a subset of the data in terms of a specific pattern language; on the other hand, by defining an interestingness

measure, we can measure the distinctness of patterns among different subgroups. However, if the target of interest consists of multiple variables with complex interactive relations, further paradigms are required, e.g. dependency modeling. Here we step into the field of EMM (Duivesteijn et al., 2016) .

EMM can be seen as an extension of SD: instead of a single target, EMM typically selects multiple target columns. A specific kind of *interaction* between these targets is captured by the definition of a *model class*. EMM finds a subgroup to be interesting when this interaction is exceptional, as captured by the definition of a *quality measure*.

Model Class In order to describe the characteristics of a local pattern, we need to choose a model class to represent the interactive relations between target variables in the associated subgroup. The model could be a statistic similar to what we done in subgroup discovery for a single target, e.g., mean and variance for multiple variables. It also can be a function that describes the dependency between multiple variables, e.g. modeling one target variable as a linear combination of the other target variables,

$$l_m = \mathbf{w}^\intercal \mathbf{l}_{\mathbf{1:m-1}},$$

where $\mathbf{w}$ represents a vector of parameters $\mathbf{w} \in \mathbb{R}^{m-1}$. The associated hypothesis space $\mathcal{W} = R^{m-1}$ is the model class we care about. Depending on what kinds of interactive relations we are interested in and the tasks we care about, different model classes could be involved. For instance, we can employ Bayesian networks (Duivesteijn et al., 2010) as the model class to investigate the mutual interactions for conditional dependence relations; we can employ a classification model (Duivesteijn and Thaele, 2014) to investigate the performance of classifiers across different subgroups. In addition, model classes can help us to capture the interestingness of subgroups on dataset with more complex data structures. For example, first-order Markov chains have been introduced as a model class for sequential data (Lemmerich et al., 2016).

Quality Measure Defining a model class allows us to describe specific characteristics in subgroups. In order to capture the interestingness, we need to define evaluation based on those characteristics, which is embodied by the quality measure:

Definition 1.1.2 A quality measure is a function $\varphi : \mathcal{D} \to \mathbb{R}$ that assigns a numeric value to a description D. Occasionally, we use $\varphi(S)$ to refer to the quality of the induced subgroup: $\varphi(S_D) = \varphi(D)$.

Typically, a quality measure assesses the subgroup at hand based on some interaction on the target columns. Hence, a description and a quality measure interact through different partitions of the dataset columns; the former focuses on the descriptors, the latter focuses on the targets, and they are linked through the subgroup. According to how a quality measure takes into account with targets, we can classify them into direct and indirect categories.

Direct Quality Measure A *direct* quality measure employ statistics on the raw target values. An example is WRAcc (Lavrač et al., 1999, van Leeuwen and Knobbe, 2011) for a binary target variable:

$$\varphi_{WRAcc}(S) = \frac{|S|}{|\Omega|}(1^S - 1^\Omega), \tag{1.1}$$

where 1^S (1^Ω) represents the fraction of ones in the subgroup (whole dataset) and $|S|$ —Ω— represents the number of records covered by the subgroup (whole dataset, i.e. = N). Another example, for a numeric target variable, is the z-score (Mampaey et al., 2015):

$$\varphi_{zscore}(S) = \frac{\sqrt{|S|}}{\sigma_0}(\mu - \mu_0), \tag{1.2}$$

where μ_0, σ_0 represent the mean and standard deviation of the single target variable in the whole dataset, μ, $\sqrt{|S|}$ represent the mean of the single target variable and the square root of number of records covered in the subgroup. Though these quality measures are defined for tabular data, some can also be adjusted for complex data structures. For instance, WRAcc was adjusted to evaluate characteristics in subgraphs (Bendimerad et al., 2016). The advantage of a direct quality measure is that the exceptionality can be easily computed and intuitively represented. The disadvantages are two-fold: on the one hand, when the dimension of the space for target variables is very high, it is difficult to compute the statistics directly; on the other hand, such statistics may not reflect the dissimilarity between distribution of targets in subgroups properly.

Indirect Quality Measure An *indirect* quality measure derives evaluations on the parameter space to compare how the models are dissimilar from each other. For instance, when two numerical columns are selected as the targets, we can consider Pearson's correlation ρ as the model class. Quality measures for this model class could be ρ itself (to find subgroups on which the target correlation is unusually high), $-\rho$ (to find subgroups with unusually strongly

negative target correlation), $|\rho|$ (to find subgroup with unusually strong positive or negative target correlation), or $-|\rho|$ (to find subgroups with unusually weak target correlation). Hence, the model class fixes the type of target interaction in which we are interested, and the quality measure fixes what, within this type of interaction, we find interesting.

Another kind of indirect quality measure is comparing the performance of the associated model in subgroups and in the whole data. This kind of exceptionality is performance based. For instance Average (Sub-)-Ranking Loss (Duivesteijn and Thaele, 2014) is proposed for testing how well the prediction of a soft classifier and the ground truth are aligned:

$$\begin{aligned}\varphi_{rasl}(S) =& \frac{\sum_{i=1}^{|S|} \mathbb{1}\{b^i = 1\} \cdot PEN_i^{|S|}(S)}{\sum_{i=1}^{N} \mathbb{1}\{b^i = 1\}} \\ &- \frac{\sum_{i=1}^{N} \mathbb{1}\{b^i = 1\} \cdot PEN_i^{N}(\Omega)}{\sum_{i=1}^{N} \mathbb{1}\{b^i = 1\}},\end{aligned} \tag{1.3}$$

$$\begin{aligned}PEN_i^{\Omega} = & \sum_{j=i+1}^{N} \mathbb{1}\{b^j = 0 \wedge r^j > r^i\} \\ &+ \frac{1}{2} \sum_{j=i+1}^{N} \mathbb{1}\{b^j = 0 \wedge r^j = r^i\},\end{aligned} \tag{1.4}$$

where b represents the binary ground truth label and r represents real-value predictions of the classifier. Lower penalty terms indicate better representation of ground truth by predictions. Lower quality values indicates less exceptionality.

Search Strategy Description language, model class and quality measure enable us to compute the interestingness of a subset from datasets. However, we still need to search in the description space in order to find the most interesting subgroups. The combination and conjunction of descriptions will lead to the pattern explosion problem (Meeng et al., 2014), hence, a smart search strategy would allow our algorithms to adapt to large datasets. The first principle to construct such a smart search strategy is the trade-off between information loss and search efficiency. The state-of-the-art consists of three ways for search strategy: exhaustive search (Atzmueller and Puppe, 2006), heuristic search (Bosc et al., 2018) and sampling search (Boley et al., 2011). Exhaustive search algorithms minimize the information loss by trying to enumerate

all the possible patterns in the search space. However, it is unfeasible to do such an exhaustive enumeration in large datasets and when the search space is infinite, e.g. continuous numeric space. Heuristic search algorithms focus on balancing the trade-off between exploration and exploitation, which allows them to be scaled to large datasets (Mampaey et al., 2012). The disadvantages of heuristic algorithms are that there is no guarantee on how the highest interestingness score is approximated and how far we are from that score. Sampling search algorithms focus on simulating a distribution of interestingness scores with respect to the support of pattern space. There are mainly two sampling strategies: input space sampling and output space sampling. The former focuses on sampling from the records of the datasets to construct the patterns of interest (Toivonen, 1996), the latter focuses on sampling from the pattern / description space directly (Al Hasan and Zaki, 2009). The limitation of sampling search algorithm is that only specific interestingness can be considered, e.g. dense neighborhood patterns with a dense measure (Giacometti and Soulet, 2018).

In this research, considering subgroups select subsets of the dataset at hand, and many such subsets exist in large amounts of data, we need to employ a search strategy to ensure that we find good results in a reasonable amount of time. Hence, we only focus on heuristic search algorithms for the practical application. Beam search is chosen as the main search algorithm. For instance, we consider the *beam search* algorithm as outlined in (Duivesteijn et al., 2016, Algorithm 1). This algorithm makes a trade-off between a pure greedy search which is likely to converge to a local optimal solution, and an exhaustive search for which it is very difficult to find the global optimum within limited time for the large scale datasets. Beam search selects candidate subgroups in a level-wise manner, by imposing a single condition on a single attribute at each step of the search. In subsequent steps, candidates with high qualities are *refined*, by attempting to extend each of these candidates with all possible additional single conditions on a single attribute, and evaluating the results. Rather than the purely greedy approach which would refine the single most promising candidate at each step, beam search refines a fixed number w (the *beam width*) of most promising candidates at each step. Larger w encourages the algorithm to explore more possibilities to escape local optima, which would take longer time. An additional parameter of beam search is the number d (the *search depth*), which sets an upper limit to the number of steps in the search process. Hence, by design, any subgroup resulting from a beam search procedure must be defined as a conjunction of at most d conditions on single attributes. Larger d implies more complex descriptions and smaller d enables the subgroups to

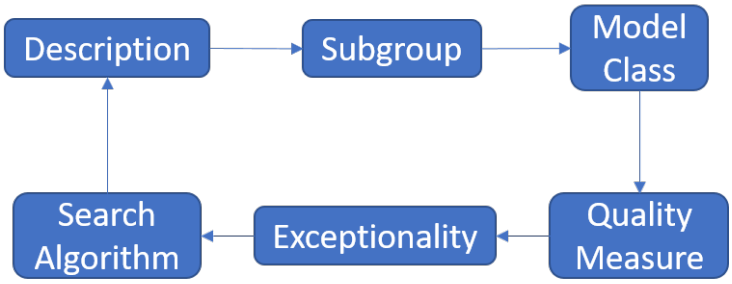


Figure 1.1: General process of EMM.

be easily interpreted, but has a more limited selectivity, fewer subgroups can be delimited.

1.2 Motivation

In the previous section, we have introduced the fundamental components of EMM. The standard research problem of EMM can be formulated as:

Problem 1.2.1 Given a dataset Ω, a description language $\mathcal{D}$, a model class Φ and a quality measure φ, our task is to find a collection of Q descriptions $h = \{D_1, \ldots, D_q\}$, such that $\forall D' \in \mathcal{D} \setminus h$, $\varphi\big(\Phi(S_{D'})\big) < \varphi\big(\Phi(S_D)\big), \forall D \in h$.

In Figure 1.1, we demonstrate the general process for solving this problem under the standard EMM framework. By defining a description language, we are able to formulate subgroups in terms of attribute variables. Coherent records covered by the same description are employed to learn the specific model regarding to the model class. Then a quality measure function is used to derive a real-valued score for the performance of the model on each subgroup. Finally, a search algorithm is applied to find the top-Q most exceptional subgroups guided by quality scores. With this framework, we can fulfill specific tasks by properly choosing the model class, quality measure and/or search algorithm. As we introduced in the previous section, one of the most important tasks is to find interesting patterns in particular form of representations from a given dataset. Nevertheless, there is one important question that needs to be answered: how certain are we about the quality score computed in this process with the dataset at hand? The other form of this question can be formed as, how to capture the uncertainty in dependency modeling with observational

data. Because models trained on limited data would have the over-confidence problem (Guo et al., 2017). Properly solving this problem is non-trivial. It can help people avoid black-box learning without knowing what indeed happens behind the number and provide fairness, accountability and transparency to the machine learning models. In particular, properly capturing the uncertainty can prevent us from being misled by false discoveries. There are three sources of uncertainty in the EMM process:

- dependency between attribute variables and target variables;
- dependency between target variables;
- dependency between quality scores.

Below we give an example to show how dependency between attributes and targets could lead to false discoveries, if our algorithms ignore the underlying causal mechanisms.

Example 1 A National Supported Work Program studies the employment status (Y) conditioning on the status of job training (X) (LaLonde, 1986, Smith and Todd, 2005). Age (Z_1), educational level (Z_2), and wealth (Z_3) of each individual are measured. Wealth affects the propensity that a person chooses to join the job training. Educational level and age jointly affect the employment status. Social economic situation (U_1), which is not measured, affects both wealth and job training. We assume the data generating process is:

$$
\begin{aligned}
&u_1 \sim \mathcal{N}(0,1),\\
&z_1 \sim \mathcal{N}(30,8),\\
&z_2 \sim \mathcal{U}(0,10),\\
&z_3|u_1 \sim (u_1 + n_3),\\
&n_3 \sim \mathcal{N}(10000,3000),\\
&x|z_3,u_1 \sim Bernoulli\Big(\sigma(w^T z_3 + u_1)\Big),\\
&w \sim \mathcal{N}(0,1\mathrm{e}{-5}),\\
\\
f_1(\mathbf{z}) &= 0.008 \cdot z_1 - 0.1 \cdot z_2,\\
f_2(\mathbf{z}) &= 0.005 \cdot z_1 - 0.3 \cdot z_2,\\
y|x,z_1,z_2 &\sim Bernoulli\Big(\sigma\big(x \cdot f_1(\mathbf{z}) + (1-x)\cdot f_2(\mathbf{z})\big)\Big),
\end{aligned}
\tag{1.5}
$$

where σ is the sigmoid function. We sample ten thousand records following this generating process. Then we hide U_1 from the synthetic data so that the

Table 1.1: Top 5 subgroups in example 1. Higher $\varphi_{\hat{Y}|X}(D)$ means more exceptional.

D	$\varphi_{\hat{Y}\|X}(D)$	$\frac{\|D\|}{N}$
$Z_2 \leq 4 \wedge Z_3 > 6305$	0.7114	.383
$Z_2 \leq 4 \wedge Z_1 > 8$	0.7097	.426
$Z_2 \leq 4 \wedge Z_3 < 14467$	0.7086	.400
$Z_2 \leq 4 \wedge Z_1 \leq 57$	0.7072	.428
$Z_2 \leq 2 \wedge Z_3 \leq 20242$	0.7067	.315

algorithms can only observe $P(Z_1, Z_2, Z_3, X, Y)$. We propose to investigate the quantity of interest $P(y = 1|x = 1)$ within and without subgroups, e.g. we can fit a Logistic regression model: $P(\hat{Y}|X;\theta) = Bernoulli\big(\sigma(\theta^\top X)\big)$ with observed data. After that, for each description, we can select the associated subsets and fit the model again. The quality is measured by comparing θ within and without subgroups. The larger the returned value is, the more exceptional that subgroup is. Then we apply beam search for the search process. The results are shown in Table 1.1. We can see that variable Z_3 is highly related to the exceptional performance in subgroups. However, Equation 1.5 indicates that Z_3 is independent of the quantity $P(\hat{Y}|X)$. This contradiction shows that the quality measure or search algorithm might be misled by the spurious associations between Z_3 and the quantity of interest. We will show in the remaining sections how to properly tackle this problem systematically. First, we need to encode the data generating process with a graph language.

Note that under the settings of this example, the exceptionality of subgroups regarding to $\{z_1, z_2, z_3\}$ and $\{z_1, z_2\}$ are observational equivalent. This means that only applying a model class and quality measure that are oblivious to the true generating process would lead to possible false discoveries. We will discuss how to overcome this problem by capturing the uncertainty properly in the following chapters.

1.3 Contributions

This dissertation studies the problem of uncertainty in EMM and explores various solutions in different application scenarios. The main questions that are answered include:

- How to capture the uncertainty in the multi-modal dependencies between target variables? How to measure the exceptionality of subgroups

in a multi-modal behavior?
- How to capture the uncertainty in the dependency between target variables and derive a proper measure to find the significantly exceptional subgroups? How to apply the solution of this question to downstream tasks, e.g. discovering exceptional educational behavior and validating the fairness of machine learning models?
- How to capture the uncertainty in the causal dependencies between treatments and outcome under the circumstance of selection/confounding bias?
- How to capture the uncertainty in the Local Causal Dependencies between a subgroup and the quantity of interest associated with that subgroup? How can Local Causal Dependencies help us to capture the exceptional subgroups?

To tackle these questions, we define new problems and develop methods, algorithms for solutions, which lead to the following contributions:

Chapter 2: uncertainty in multi-modal dependency modeling In this chapter, we introduce the problem of discovering exceptional subgroups considering the multi-modal dependencies between target variables. In order to capture such multi-modal dependencies, we propose to explicitly simulate the underlying data generating process by building Bayesian non-parametric models. Based on the learned Bayesian non-parametric models, we propose to quantify the exceptionality by comparing the posterior distributions of model parameters in subgroups and the whole data. Finally, the methods and results are applied to spatio-temporal behavior analysis which allows us to detect exceptional subgroups considering their spatial, time and text activity on social media like twitter (Du et al., 2020b).

Chapter 3: uncertainty in dependency modeling In this chapter, we consider to capture the exceptionality in subgroups with heterogeneous and high-dimensional interactions between target variables. We introduce the dependency modeling methods by point estimate, and propose defining quality measures based on learning with specific model classes. Then we propose hypothesis testing to capture the uncertainty between attributes and targets to validate the significance of the measured qualities. Finally, the methods and results are applied to practical applications to study how EMM can help people to understand the educational behavior (Du et al., 2018) and the fairness in network representation models (Du et al., 2020c).

Chapter 4: uncertainty in causal dependency In this chapter, we study a specific dependency between variables: causal dependency. Instead of computing correlation dependency by conditional probability, the calculation of causal dependency requires to model the intervention process on variables. The intervention process can be represented by the do-operator (Pearl, 2009). Due to the confounding / selection bias, it is difficult to compute the causal dependency from observational data. We study this problem from the view of counterfactual prediction with the Potential Outcome framework (PO) (Rubin, 2005). A neural network framework employing adversarial balanced representation learning is proposed to estimate the causal dependency (Du et al., 2019).

Chapter 5: uncertainty in Local Causal Dependency In this chapter, we study a the problem of EMM by considering the causal dependency between attributes and the quantity of interest. A computational graph, D-graph, is proposed to capture this specific dependency by using the structural causal model (Du et al., 2020a). We show how this new defined problem can prevent us from being misled by false discoveries comparing with tradition EMM. This method can help us to understand why a given model performs differently in subgroups defined in terms of attributes.

2

Uncertainty in Multi-modal Dependency Modeling

"I wanted certainty in the kind of way in which people want religious faith. I was continually reminded of the fable about the elephant and the tortoise. Having constructed an elephant upon which the mathematical world could rest, I found the elephant tottering, and proceeded to construct a tortoise to keep the elephant from falling."

Portraits from Memory and Other Essays,
Bertrand Russell, 1956.

2.1 Introduction

In this chapter, we propose to discuss the uncertainty in EMM under the condition of multi-modal dependency. For instance, collective social media provides a vast amount of geo-tagged social posts, which contain various records on spatio-temporal behavior. Modeling spatio-temporal behavior on collective social media is an important task for applications like tourism recommendation, location prediction and urban planning. Properly accomplishing this task requires a model that allows for diverse behavioral patterns on each of the three aspects: spatial location, time, and text. Traditional methods in SD / EMM consider the distribution of a single target, or single model class for the evaluation of exceptional behavior in subgroups. However, when the targets of interest consist of multi-variables with multi-modal behavior, existing quality measures and model classes may not be able to capture the real exceptionality. We propose to build a systematic method to solve this problem with a specific case: how to find representative subgroups of social posts, for which the spatio-temporal behavioral patterns are substantially different from the behav-

ioral patterns in the whole dataset?

The challenges for solving this problem are two-folds: on the one hand, we need to develop a new model class that can capture the multi-modal behavior from observational data; on the other hand, a new quality measure is required to capture the exceptionality of multi-modal behavior with limited data records. With limited data records, the training process of a model may not return the optimum hypothesis, which could bring more uncertainty for the evaluation of exceptionality. Point estimate methods for single modal model class could not provide enough confidence for the exceptionality. We propose to quantify the exceptionality of a subgroup with regard to the data generating process.

2.2 Motivation

Popular social media platforms such as Twitter and Instagram have millions of users who share their photos, stories and geo-locations. This allows the collective social media to reflect diverse human behavioral patterns. The behavioral patterns in social posts are represented by joint distributions of spatial locations, time, and word topics (Hong et al., 2012). Specific deviations across any combination of these three distributions can indicate interesting, exceptional behavior of the population; one can for instance see such deviations surrounding large events, such as sports games and concerts (Zheng et al., 2018). In this chapter, instead of social posts for individuals, we are interested in finding social posts for subgroups restricted by descriptions, for which the behavioral patterns are substantially different compared to the behavioral patterns in the whole dataset. Discovering and understanding these behavioral patterns on collective social media is a task of predominant importance, since properly accomplishing this task can benefit applications such as tourism recommendation, location prediction, and urban planning (Kim et al., 2016).

To contribute to this behavioral understanding, instead of finding outlying social posts far from the main activity areas, we are looking for exceptional subgroups: coherent subsets for which we can formulate concise descriptions in terms of conditions on attributes of the data (Herrera et al., 2011, Atzmueller, 2015), e.g., 'Age $< 25 \wedge$ gender = Female'. The most challenging problems for finding exceptional subgroups are: how to model the spatio-temporal behavior and quantify the exceptionality of the subgroups? Before proposing the solution, we discuss the challenges which need to be overcome at first:

Spatio-temporal modeling. Difficulties stem from two aspects. On the one hand, unlike modeling behavior of individuals, where the records are grouped by certain subjects (Yuan et al., 2017), in our problem setting, the candidate subgroups are apriori unknown. We cannot model the spatio-temporal behavior of all the subgroups either, because of the pattern explosion problem (Meeng et al., 2014). This means that we cannot directly model the global distribution of behavioral patterns over the whole dataset. On the other hand, collective social activities typically contain uncertain spatial, temporal, and text information on diverse scales (Jankowiak and Gomez-Rodriguez, 2017). To properly overcome these challenges, we need a model that can handle the diverse, uncertain, large scale, and high-dimensional information in collective social posts and induce the global distribution of behavioral patterns in the whole dataset.

Exceptionality evaluation. Our aim is to identify exceptional behavioral patterns of social posts in subgroups. The general method would be to learn the joint distributions of spatial locations, time, and texts empirically by probability mass (Giannotti et al., 2016), followed by comparing the distributions in subgroups with the global distributions in the whole dataset. However, this method is not applicable for the research problem of this chapter. The reasons are two-fold. On the one hand, given limited records, we cannot be confident whether a subgroup is exceptional or not in long term behavior only by comparing the empirical distributions. On the other hand, because of the uncertainty and diversity of social posts in collective social media, it is difficult to simply assume a distribution for the behavioral patterns and build a null hypothesis to test (Hooi et al., 2016).

2.3 Contributions

To overcome these challenges, we propose **BNPM**: a **B**ayesian **n**on-**p**arametric **m**odel for spatio-temporal behavior modeling on the subgroup level. In BNPM, we randomly sample arbitrarily large numbers of subgroups as the training samples in order to estimate the global behavior. We employ a Chinese Restaurant Process (CRP) to gather those randomly sampled subgroups into several components. In this process, the behavioral pattern of each subgroup is assumed to follow a prior distribution. Subgroups in one CRP component are allowed to have variations in distribution, but similar kinds of behavior ought to aggregate within every single component. Hence, the CRP

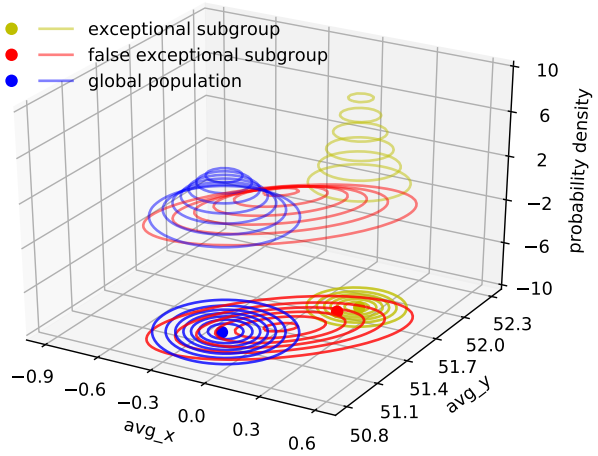


Figure 2.1: Comparison between Bayesian posterior distribution and point estimate. Contours represent the distribution of μ (mean of spatial locations) following a multivariate Gaussian distribution; solid points represent point estimates of μ.

model allows for modeling multiple types of normal behavior to occur simultaneously, which more accurately represents real life than if we assume one monolithic kind of normal behavior. We estimate the global distribution of behavioral patterns in the whole dataset by the mixture of behavioral patterns with mixture coefficients of the components (cf. Equation (2.19)). Specifically, for each given subgroup, we can calculate its posterior distribution with the learned BNPM, according to the information of spatial locations, time, and texts. The exceptionality score of the given subgroup is derived by computing the distance between the posterior distribution and the global distribution. We employ a variant of weighted KL-divergence (van Leeuwen and Knobbe, 2012) for multi-variate distribution (Soch and Allefeld, 2016), to calculate the distance between the posterior distribution of the subgroup and the global distribution. Finally, we aggregate the exceptionality scores in the aspects of spatial locations, time, and texts as the final exceptionality score of the candidate subgroup.

In Figure 2.1, we present an artificial example to show the advantage of our method. From the perspective of a point estimate, both the red and the yel-

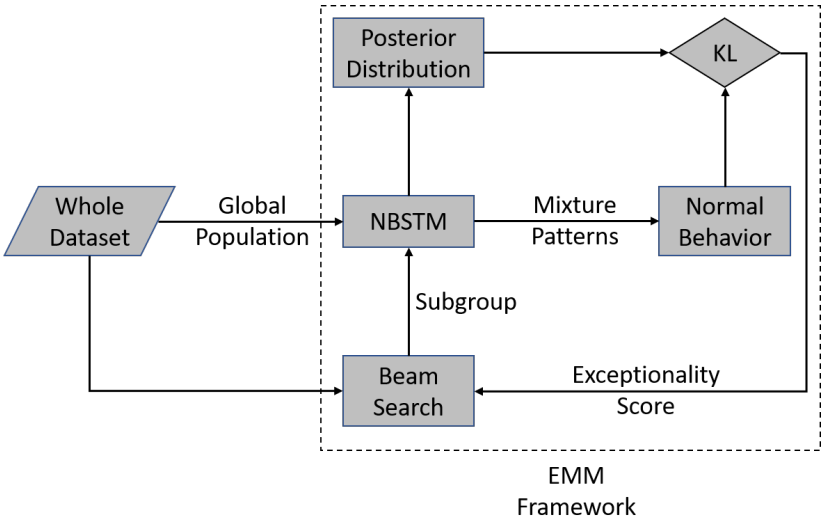


Figure 2.2: Methodological pipeline involving BNPM.

low subgroups are exceptional compared with the global population (in blue). However, from the perspective of Bayesian posterior distribution, the yellow one is much more suspicious than the red one. The reason is that the point estimate uses limited data to estimate the behavioral pattern, which might lead to biased results. The Bayesian non-parametric method evaluates the exceptionality of behavioral patterns by comparing the posterior distribution with the global distribution, which can help us effectively find exceptional behavioral patterns and prevent false discoveries.

The training process of our model includes two iteration steps: assigning subgroups into components and updating hyper-parameters for the components. These two processes influence each other iteratively. We integrate these two steps with the collapsed Gibbs sampling (Porteous et al., 2008) algorithm. Having learned the well-trained model over the whole dataset, we can calculate the posterior distribution for any subgroup across the location distribution, time distribution, and text distribution. This allows us to employ EMM to automatically discover subgroups with exceptional spatio-temporal behavior. The whole process of our method is shown in Figure 2.2. To demonstrate the effectiveness and scalability of our method, we validate our model by conducting experiments on four real-world datasets from New York, London, Tokyo, and Shenzhen. The resulting subgroups illustrate the versatility of the method. In London, our method discovers the spatially coherent subgroup of people at-

tending a specific football match. In Tokyo, it discovers a subgroup of people frequenting three locations in a specific ward: two touristic attractions and a station where trains leave for a third touristic attraction (identified by analyzing the texts of the tweets) which is located relatively far away. The combination of spatio-temporal behavior and tweet text behavior can benefit the uncovering of such a subgroup, which is where the added value of our method lies. Finally, in another ward of Tokyo, two subgroups separate the professionals and the tourists by their combined spatio-temporal and tweet text behavior. In summary:

- We introduce **BNPM**: a Bayesian Non-Parametric Model for spatio-temporal behavior modeling on the subgroup level. BNPM can handle diverse, uncertain, large scale and multi-modal information in collective spatio-temporal data.
- We define a new evaluation method for EMM. The global distribution is generated by the mixture of behavioral patterns in BNPM. By comparing the posterior distribution of a candidate subgroup with the global distribution, we can quantify the exceptionality of subgroups.
- We conduct various experiments on four real-world datasets. The results show that our method is effective and efficient for finding exceptional social posts on the subgroup level.

2.4 Related Work

Exceptional spatio-temporal behavior mining on the subgroup level is related to three fields: anomaly detection (Chandola et al., 2009), EMM (Duivesteijn et al., 2016) in the aspect of exceptionality metric; and spatio-temporal modeling (Atluri et al., 2017) in the aspect of behavior modeling.

Anomaly Detection Anomaly detection is highly explored in online ratings (Hooi et al., 2016), reviews (Xie et al., 2012), and social network analysis (Shin et al., 2017). In order to detect collective anomalies on spatio-temporal datasets with different distributions, densities and scales, researchers have proposed a multi-source topic model for spatio-temporal modeling (Wu et al., 2017, Zheng et al., 2015). Methods such as classification, statistical, and regression models are used for modeling user behavior to discover anomaly patterns (Shipmon et al., 2017).

Unlike anomaly detection, there is no labeled data for identifying anoma-

lies in EMM. This means that standard supervised learning cannot be used directly for this task. The exceptional subgroups are identified by comparing the performance of the model in subgroups with the performance of the model in the whole dataset, for which the subgroups are restricted by the descriptive variables (Duivesteijn et al., 2016). The whole process of EMM lies in the field of knowledge discovery. This formulates the main difference between the research of anomaly detection and EMM.

Exceptional Model Mining Though existing model classes can handle all kinds of targets, most cannot model spatio-temporal behavior, which contains geo-spatial coordinates and timestamps. Lemmerich et al. (2016) introduce first-order Markov chains as a model class for sequence data, which can be used for mining exceptional transition behavior. Bendimerad et al. (2016) employ weighted relative accuracy to evaluate characteristics in subgraphs of urban regions. However, they do not consider the text information, especially the word topics. This information integration is the added value of our model. In order to properly handle the noise inherent to spatial and temporal data and prevent false positives, we introduce a quality measure under the Bayesian framework.

Spatio-Temporal Modeling There is a vast amount of literature about spatio-temporal data mining (Atluri et al., 2017, Lane et al., 2014, Wang et al., 2011, Yuan et al., 2017). Most work focuses on modeling mobility patterns of individuals or groups aiming at location prediction or period discovery. The basic assumption is that individuals or groups might have a regular activity area, which indicates the inner similarity of social and geographic closeness (Cranshaw et al., 2010). Piatkowski et al. (2013) present a graphical model designed for efficient probabilistic modeling of spatio-temporal data, which can keep the accuracy as well as efficiency. Knauf et al. (2016) propose a spatio-temporal kernel for multi-object scenarios. A branch of research focuses on visual analytics for spatio-temporal modeling (Zheng et al., 2016). Interactive and human-guided methods are employed to discover the behavioral patterns and understand the heterogeneous information in the urban data (Puolamäki et al., 2016, Chen et al., 2018c). The differences between our work and the work before are two-fold. On the one hand, the collective social posts on the subgroup level in our research is constrained by the descriptions, which distinguishes our work from others such as twitter stream clustering or user clustering (Chierichetti et al., 2014). On the other hand, the exceptional subgroups and the components of behavioral distributions are unobserved from the

datasets, which means that we have to establish a model for the modeling of global distribution of behavioral pattern as well as discovering the exceptional subgroups comparing with this global distribution.

2.5 Methodology

2.5.1 Preliminaries

Assume a dataset Ω: a bag of m records $r \in \Omega$ of the form:

$$r = (a_1, \ldots a_s, b_1, \ldots b_u),$$

where s and u are positive integers. We call $a_1, \ldots, a_s$ the *descriptive attributes* or *descriptors* of r, and $b_1, \ldots, b_u$ the *target attributes* or *targets* of r. The descriptive attributes are taken from an unrestricted domain $\mathcal{A}$. Mathematically, we define descriptions as functions $D : \mathcal{A} \to \{0, 1\}$. A description D covers a record r^j if and only if $D(a_1^j, \cdots, a_s^j) = 1$.

A Chinese Restaurant Process (CRP) (Blei et al., 2010) is a distribution on partitions of integers obtained by imagining a process by which $n-1$ customers sit down in a Chinese restaurant with an infinite number of tables with infinite capacity. Whenever a new customer arrives, customer n, she can either choose an existing table k with n_k seated customers or sit at an empty table, following distribution:

$$p(\text{existing table } k \mid \text{previous customers}) = \frac{n_k}{n-1+\alpha},$$
$$p(\text{new table} \mid \text{previous customers}) = \frac{\alpha}{n-1+\alpha}.$$

In each step a new table is created with non-zero probability, which allows this process to adapt to increasing complexity of the data.

2.5.2 Subgroup-Level Spatio-Temporal Modeling (BNPM)

We consider the spatio-temporal patterns of geo-tagged social posts on the level of subgroups restricted by descriptive attributes. For notational purposes, we ignore that these subgroups need to be generated somehow; instead, we assume that some process has delivered us a list of subgroups, indexed $i = 1, \ldots, n$, where subgroup i has d_i posts, indexed by $j = 1, \ldots, d_i$. The posts in subgroup i are denoted by the variables $r_{ij} \in \{1, 2, \ldots, m\}$; posts may

Table 2.1: Notations used in the chapter.

Notation	Description
n	Number of subgroups
m	Number of geo-tagged social media posts
d_i	Number of posts belongs to subgroup i
D	Description of a subgroup
r_{ij}	Social media post j in subgroup i
$l_{ij} = (x, y)$	Spatial location of post j in subgroup i
$t_{ij} = t$	Time of post j in subgroup i
$w_{ij} = \{w_1, \ldots, w_q\}$	Texts of post j in subgroup i
n_k	Number of subgroups in component k
z_i	Component assignment of subgroup i
K	Number of components
V	Vocabulary of the whole words
α	Concentration parameter of CRP
β_k	Probability to choose component k
μ_i, Σ_i	Mean and covariance of spatial locations in subgroup i
υ_i, σ_i	Mean and variance of time in subgroup i
θ_i	Word distribution for posts in subgroup i
$\mu_{0z_i}, \lambda_{z_i}, W_{z_i}, \nu_{z_i}$	Normal-Inverse-Wishart ($\mathcal{NIW}$) prior for μ_i, Σ_i
$\upsilon_{0z_i}, \kappa_{z_i}, \rho_{z_i}, \psi_{z_i}$	Normal-Gamma ($\mathcal{NG}$) prior for υ_i, σ_i^2
θ_{0z_i}	Dirichlet prior for θ_i

belong to multiple subgroups. Each post is a 3-tuple r_{ij} = (l_{ij},t_{ij},w_{ij}), where l_{ij} = (x,y), t_{ij} = t and $w_{ij} = \{w_1, \ldots, w_q\}$ represent the spatial location, time, and a bag of words in a geo-tagged post. Table 2.1 lists the notations used in the rest of this chapter. We now propose the problem of discovering subgroups with exceptional spatio-temporal behavior as follows:

Problem 2.5.1 (Discovering subgroups with exceptional spatio-temporal behavior) Given a dataset of geo-tagged social posts Ω, descriptive attributes taken from $\mathcal{A}$, descriptions $D : \mathcal{A} \rightarrow \{0, 1\}$, and a quality measure φ, our aim is to find a bag of subgroups $\{S_{D_1}, \cdots, S_{D_q}\}$, where $\forall D' \in \mathcal{D} \setminus \{D_1, \cdots, D_q\}, \forall D \in \{D_1, \cdots, D_q\}, \varphi(D') \leq \varphi(D)$.

The main challenge for this problem is the subgroup selection process with regard to the exceptionality compared with the global population. To accomplish this task, we need a spatio-temporal model on the subgroup level, to

model the behavioral patterns in the global population and subgroups.

The Bayesian Non-Parametric Model Several intuitions underpin our model:

1. The behavioral patterns of subgroups over the whole dataset can be captured by several components. Each component follows a single triplet of prior distributions: of spatial locations, time, and word topics. We assume that the social posts are generated by the mixtures of components with mixture coefficients, but the number of components and the mixture coefficients are unobserved from the dataset.
2. Despite following the same *prior* distribution, subgroups within the same component need not have the *same* distributions of spatial locations, posting time, and texts.
3. Social posts are distributed in spatial regions, with time ranges as well as word topics. These distributions indicate the spatio-temporal behavioral patterns of subgroups. The spatio-temporal behavioral pattern varies according to the center and scale of the region and time, as well as the word topics.

Based on these intuitions, we assume that subgroups and social posts are governed by a generative model. This model for spatio-temporal behavior on the subgroup level is a mixture model in which each subgroup belongs to one of the components, in order to capture different types of behavior. Each component represents a behavioral pattern with specific prior distributions of location, time, and word topics. The spatial location associated to each geo-tagged post is drawn from a multivariate Gaussian distribution, as suggested by Gonzalez et al. (2008):

$$l = (x, y) \sim \mathcal{N}(\mathbf{l}|\mu, \Sigma).$$

For each component, we assume that a Normal-Inverse-Wishart (NIW) distribution is the prior distribution that governs the generation of means and covariance matrices (μ, Σ) for spatial locations, as suggested by Yuan et al. (2017):

$$(\mu, \Sigma) \sim \mathcal{NIW}(\mu, \mathbf{\Sigma}|\mu_0, \lambda, W, \nu).$$

Similarly, we can write down the generative process of time t from a univariate Gaussian distribution, as suggested by Cho et al. (2011), as:

$$t \sim \mathcal{N}(\mathbf{t}|\upsilon, \sigma^2), \tag{2.1}$$

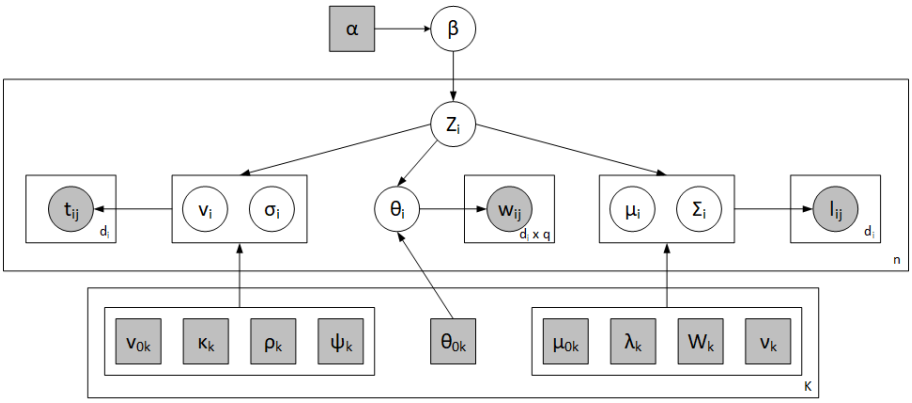


Figure 2.3: Graphical model representing subgroups with locations, time and texts of posts. Shaded rectangles are hyper-parameters, blank circles are latent variables and shaded circles are observations.

where the mean υ and variance σ are drawn from a Normal-Gamma prior distribution, as suggested by Yuan et al. (2017):

$$(\upsilon, \sigma) \sim \mathcal{NG}(\upsilon, \sigma | \upsilon_0, \kappa, \rho, \psi). \tag{2.2}$$

Each word w in $\{w_1, \ldots, w_q\}$ is drawn from a multinomial distribution, as suggested by Jankowiak and Gomez-Rodriguez (2017):

$$w \sim Mult(\theta), \tag{2.3}$$

where θ is a distribution that represents proportions of words in vocabulary V, which depends on the Dirichlet prior θ_0 (Jankowiak and Gomez-Rodriguez, 2017):

$$\theta \sim \text{Dirichlet}(\theta_0). \tag{2.4}$$

By construction, the proposed generative model gathers the subgroups into several components, which raises the question of choosing the number of components. If we set the number too high, spatio-temporal behavioral patterns of subgroups may vary too much, which will impede proper evaluation of behavior exceptionality. Conversely, if we set the number too low, exceptional subgroups may be mixed with normal subgroups, which will lead to false positive errors. This is where we employ the Chinese Restaurant Process (cf. Section 2.5.1). The full generative process (cf. Figure 2.3) can be summarized as follows:

1. Set the number of components K $\leftarrow 0$

2. For $i = 1, \ldots, n$:
   (a) Assign subgroup i to an existing component $k \in \{1, \ldots, K\}$ with probability $\beta_k = \frac{n_k}{i-1+\alpha}$, or to a new component $k = K+1$ with probability $\frac{\alpha}{i-1+\alpha}$.
   (b) Draw $(\mu_i, \Sigma_i)|z_i = k \sim \mathcal{NIW}(\mu_{0k}, \lambda_k, W_k, \nu_k)$.
   (c) Draw $(\upsilon_i, \sigma_i)|z_i = k \sim \mathcal{NG}(\upsilon_{0k}, \kappa_k, \rho_k, \psi_k)$.
   (d) Draw $\theta_i|z_i = k \sim$ Dirichlet(θ_{0k}).
   (e) For j = 1,. . . ,d_i:
      i. Draw $l_{ij} \sim \mathcal{N}(\mathbf{l}|\mu_i, \Sigma_i)$.
      ii. Draw $t_{ij} \sim \mathcal{N}(\mathbf{t}|\upsilon_i, \sigma_i^2)$.
      iii. Draw each $w_{ijq} \in \{w_1, \ldots, w_q\} \sim Mult(\mathbf{w}|\theta_i)$.
   (f) Update hyper-parameters in component k.

Inference Method As illustrated above, to conduct the whole generating process, we need to estimate the latent variables, which cannot be observed directly from the datasets. We propose to employ collapsed Gibbs sampling to infer the latent variables in the proposed generative model efficiently (Porteous et al., 2008). Given full observation of n subgroups, the total likelihood is:

$$\begin{aligned} &P(\mathbf{l},\mathbf{t}, \mathbf{w}, \mathbf{z}|\alpha, \mu_0, \lambda, W, \nu, \upsilon_0, \kappa, \rho, \psi, \theta_0) \\ &= \int_\beta P(\mathbf{z}|\beta)P(\beta|\alpha)d\beta \cdot \int_\mu \int_{\mathbf{\Sigma}} P(\mathbf{l}|\mu, \mathbf{\Sigma})P(\mu, \mathbf{\Sigma}|\mu_\mathbf{0}, \lambda, \mathbf{W}, \nu)d\mu d\mathbf{\Sigma} \\ &\cdot \int_\mathbf{v} \int_\sigma P(\mathbf{t}|\upsilon, \sigma)P(\upsilon, \sigma|\mathbf{v_0}, \kappa, \rho, \psi)d\mu d\sigma \cdot \int_\theta P(\mathbf{w}|\theta)P(\theta|\boldsymbol{\theta_0})d\theta. \end{aligned} \tag{2.5}$$

We exploit the conjugacy between the multinomial and Dirichlet distributions, the Gaussian and Normal-Inverse-Wishart distributions, and the Gaussian and Normal-Gamma distributions. Hence, we can analytically integrate out the parameters $\beta, \mu, \mathbf{\Sigma}, \upsilon, \sigma$, and θ, and only sample the component assignments $\mathbf{z}$, which is done as follows:

$$\begin{aligned} &P(z_i = k|\mathbf{z}_{\neg\mathbf{i}}, \mathbf{l_i}, \mathbf{t_i}, \mathbf{w_i}, \alpha, \mu_{0k}, \lambda_k, W_k, \nu_k, \upsilon_{0k}, \kappa_k, \rho_k, \psi_k, \theta_{0k}) \propto \\ &P(z_i = k|\mathbf{z}_{\neg\mathbf{i}}, \alpha) \cdot P(\mathbf{l_i}|\mathbf{l}_{\neg\mathbf{i}}, \mu_{0k}, \lambda_k, W_k, \nu_k) \\ &\cdot P(\mathbf{t_i}|\mathbf{t}_{\neg\mathbf{i}}, \upsilon_{0k}, \kappa_k, \rho_k, \psi_k) \cdot P(\mathbf{w_i}|\mathbf{w}_{\neg\mathbf{i}}, \theta_{0k}). \end{aligned} \tag{2.6}$$

The first term of Formula (2.6) is governed by the CRP:

$$P(z_i = k|\mathbf{z}_{\neg\mathbf{i}}, \alpha) = \begin{cases} \frac{n_{k\neg i}}{n-1+\alpha} & \text{if } k \text{ exists,} \\ \frac{\alpha}{n-1+\alpha} & \text{if } k \text{ is new.} \end{cases} \tag{2.7}$$

The second term is the posterior predictive distribution of $\mathbf{l_i}$ in component k, excluding subgroup i. We assume that each post in subgroup i is generated equivalently, hence the second term equals:

$$\begin{aligned} &\prod_{j=1}^{d_i} p(l_{ij}|\mathbf{l_{k\neg i}}, \mu_{0k}, \lambda_k, W_k, \nu_k) \\ =&\prod_{j=1}^{d_i} \tau_{\nu_{nk}-1}\left(l_{ij}\middle|\mu_{n_k\neg i}, \frac{\lambda_{n_k}+1}{\lambda_{n_k}(\nu_{n_k}-1)}W_{n_k\neg i}\right). \end{aligned} \tag{2.8}$$

Here, $\mathbf{l_{k\neg i}}, n_{k\neg i}$ are locations, and the number thereof in component k after excluding subgroup i,

$$\begin{aligned} \mu_{n_k\neg i} &= \frac{\lambda_k\mu_{0k} + n_{k\neg i}\bar{l}_{k\neg i}}{\lambda_{n_k}}, \quad \lambda_{n_k} = \lambda_k + n_{k\neg i}, \\ W_{n_k\neg i} &= W_k + \sum_{l\in\mathbf{l_{k\neg i}}} (l-\bar{l}_{k\neg i})(l-\bar{l}_{k\neg i})^T \\ &+ \frac{\lambda_k n_{k\neg i}}{\lambda_k + n_{k\neg i}}(\bar{l}_{k\neg i}-\mu_{0k})(\bar{l}_{k\neg i}-\mu_{0k})^T, \quad \nu_{nk} = \nu_k + n_{k\neg i}. \end{aligned} \tag{2.9}$$

The posterior predictive distribution of each l_{ij} follows a bivariate Student's t-distribution (Murphy, 2007). Similarly, we can write down the posterior predictive distribution of $\mathbf{t_i}$ in the third term of Formula (2.6):

$$\prod_{j=1}^{d_i} \tau_{2\rho_{n_k}}\left(t_{ij}\middle|\upsilon_{n_k\neg i}, \frac{\psi_{n_k\neg i}(\kappa_{n_k}+1)}{\rho_{n_k}\kappa_{n_k}}\right), \quad \text{where} \tag{2.10}$$

$$\begin{aligned} \upsilon_{n_k\neg i} &= \frac{\kappa_k\mu_{0k} + n_{k\neg i}\bar{t}_{k\neg i}}{\kappa_{n_k}}, \quad \kappa_{n_k} = \kappa_k + n_{k\neg i}, \quad \rho_{nk} = \rho_k + n_{k\neg i}/2 \\ \psi_{n_k\neg i} &= \psi_k + \frac{1}{2}\sum_{t\in\mathbf{t_{k\neg i}}} (t-\bar{t}_{k\neg i})^2 + \frac{\kappa_k n_{k\neg i}(\bar{t}_{k\neg i}-\upsilon_{0k})^2}{2\kappa_{n_k}}. \end{aligned} \tag{2.11}$$

The posterior predictive distribution of each t_{ij} follows a univariate Student's t-distribution. For the fourth term of Formula (2.6), each posterior predictive

distribution of $\mathbf{w_{ij}}$ for post j in subgroup i follows a Dirichlet-multinomial distribution (Tu, 2014):

$$P(\mathbf{w_{ij}}|\theta_{0k}) = \frac{\Gamma(c_{k\neg i} + V\theta_{0k}) \prod_{w \in V} \Gamma(c_{wk\neg i} + c_{wj} + \theta_{0k})}{\Gamma(c_{k\neg i} + c_j + V\theta_{0k}) \prod_{w \in V} \Gamma(c_{wk\neg i} + \theta_{0k})}. \tag{2.12}$$

Here, $c_{k\neg i}$ is total number of words in component k so far excluding subgroup i, $c_{wk\neg i}$ is how often word w occurs in component k so far excluding subgroup i, c_j is the total number of words in post ij, and c_{wj} is how often word w occurs in post ij.

Our model assumes that each component has its own specific hyper-parameters. If we fix all the assignments of $\mathbf{z}$, we use random search for hyper-parameter optimization (Bergstra and Bengio, 2012) to choose $\mu_{0k}, \lambda_k, W_k, \nu_k, \upsilon_{0k}, \kappa_k, \rho_k, \psi_k$, and θ_{0k}. Our goal is maximizing the marginal likelihood of the data in each component (Bergstra et al., 2011):

$$\underset{(\mu_{0k}, \lambda_k, W_k, \nu_k)}{\operatorname{argmax}} \; P(\mathbf{l_k}|\mu_{0k}, \lambda_k, W_k, \nu_k), \tag{2.13}$$

$$\underset{(\upsilon_{0k}, \kappa_k, \rho_k, \psi_k)}{\operatorname{argmax}} \; P(\mathbf{t_k}|\upsilon_{0k}, \kappa_k, \rho_k, \psi_k), \tag{2.14}$$

$$\underset{\theta_{0k}}{\operatorname{argmax}} \, P(\mathbf{w_k}|\theta_{0k}). \tag{2.15}$$

Now, we can build up the two iteration processes in our inference algorithm. The one is iteratively optimizing hyper-parameters for fitting subgroups in associated components. The other is iteratively sampling component assignments to assign subgroups. These two steps influence each other: better hyper-parameter selection provides more accurate posterior predictive distribution to assign subgroups; better assignments for subgroups can provide more accurate likelihood estimation for hyper-parameter selection. We iteratively run these two steps until a maximum number of iterations is reached. See Algorithm 1 for details.

Subgroup Evaluation Method Having learned the proposed model, we need to evaluate the exceptionality of a subgroup. Behavioral patterns are gauged in terms of the location distribution, time distribution, and text distribution. As an example, we use time distribution to explain our method for exceptionality evaluation. Let $\mathbf{t_i}$ denote a vector representing the post time of collective social posts in subgroup i. Generally, people will assume a distribution for $P(t)$, e.g., $\mathcal{N}(\upsilon, \sigma)$, and use the point estimate of υ and σ as the estimated parameters of that distribution. The learned distribution is regarded

Algorithm 1 Inference algorithm for BNPM.

Initialize $\mathbf{z}$, $\mu_{0k}, \lambda_k, W_k, \nu_k, \upsilon_{0k}, \kappa_k, \rho_k, \psi_k, \theta_{0k}$;
Initialize α;
while not reach the maximum iterations **do**
    for $k = 1$ to K **do**
        Update $\mu_{0k}, \lambda_k, W_k, \nu_k$ using Formula (2.13);
        Update $\upsilon_{0k}, \kappa_k, \rho_k, \psi_k$ using Formula (2.14);
        Update θ_{0k} using Formula (2.15);
    end for
    for $i = 1$ to n **do**
        Exclude i from component z_i;
        for $k = 1$ to K **do**
            Compute $P(z_i = k|\mathbf{z}_{\neg \mathbf{i}}, \alpha)$ using Equation (2.7);
            Compute $P(\mathbf{l_i}|\mathbf{l_{k\neg i}}, \mu_{0k}, \lambda_k, W_k, \nu_k)$ using Equation (2.8);
            Compute $P(\mathbf{t_i}|\mathbf{t_{k\neg i}}, \upsilon_{0k}, \kappa_k, \rho_k, \psi_k)$ using Formula (2.10);
            Compute $P(\mathbf{w_i}|\mathbf{w_{k\neg i}}, \theta_{0k})$ using Equation (2.12);
            Compute $P(z_i = k|\mathbf{z}_{\neg \mathbf{i}}, .)$ using the preceding results;
        end for
        Compute $P(z_i = k^*|\mathbf{z}_{\neg \mathbf{i}}, \alpha)$ using Equation (2.7);
        Compute $P(\mathbf{l_i}|\mu_{0k^*}, \lambda_{k^*}, W_{k^*}, \nu_{k^*})$ using Equation (2.8);
        Compute $P(\mathbf{t_i}|\upsilon_{0k^*}, \kappa_{k^*}, \rho_{k^*}, \psi_{k^*})$ using Formula (2.10);
        Compute $P(\mathbf{w_i}|\theta_{0k^*})$ using Equation (2.12);
        Compute $P(z_i = k^*|\mathbf{z}_{\neg \mathbf{i}}, .)$ using the preceding results
        Sample k_{new} from $P(z_i|\mathbf{z}_{\neg \mathbf{i}}, .)$;
        Update component $z_i = k_{new}$;
        if $k_{new} > K$ **then**
            $K = K + 1$;
        end if
        if any component k is empty **then**
            $K = K - 1$;
        end if
    end for
end while

as an estimation about the temporal behavioral pattern of subgroup i. However, this distribution is not sufficient to represent the real behavioral pattern of subgroup i, because we cannot be confident about the behavior of that subgroup with limited records. Hence, in this chapter, instead of a point estimate for a distribution with limited data, we compute the posterior distribution as our belief about the behavioral pattern of a subgroup. For each given candidate subgroup i, we firstly estimate the component assignment z_i on this subgroup by using Formulas (2.6), (2.7), (2.8), (2.10), and (2.12). Then, with BNPM, we calculate the posterior distribution of subgroup i's location distribution, time distribution, and text distribution:

$$P(\mu, \Sigma|\mathbf{l_i}) = \mathcal{NIW}(\mu, \Sigma|\mu_{0z_i}, \lambda_{z_i}, W_{z_i}, \nu_{z_i}), \tag{2.16}$$

$$P(\upsilon, \sigma|\mathbf{t_i}) = \mathcal{NG}(\upsilon, \sigma|\upsilon_{0z_i}, \kappa_{z_i}, \rho_{z_i}, \psi_{z_i}), \tag{2.17}$$

$$P(\theta|\mathbf{w_i}) = \text{Dirichlet}(\theta|\theta_{0z_i}). \tag{2.18}$$

Here we calculate the posterior parameters the same way as Equations (2.9), (2.11), and (2.12), with the prior hyper-parameters in component z_i. Having obtained the posterior distribution, the next step is to evaluate the exceptionality. In the training process, we learn the mixture proportion of components denoted as β. The global distribution of time is governed by both components and the mixture proportion of components. We can calculate the distribution of time in the global population by Equation (2.2) as:

$$P(\upsilon, \sigma) = \sum_{k=1}^{K} \beta_k \cdot \mathcal{NG}(\upsilon, \sigma|\upsilon_{0k}, \kappa_k, \rho_k, \psi_k). \tag{2.19}$$

This distribution describes the temporal behavioral pattern averaged by the global population. Now we can compare the posterior distribution of time conditioned on a subgroup, with the global distribution of time. The more different they are, the more exceptional the subgroup is. The difference indicates how difficult it is to generate the time distribution in that subgroup under the global population. In order to quantify this difference, we employ KL-divergence as the distance measure between two distributions. For simplicity, we represent Equation (2.17) with $f(\upsilon, \sigma)$ and Equation (2.19) with $g(\upsilon, \sigma) = \sum_{k=1}^{K} \beta_k \cdot g_k(\upsilon, \sigma)$. The exceptionality score of a given subgroup i

in the time aspect is:

$$\begin{aligned}\varphi_{t_i} =& \frac{d_i}{m} D_{KL}(f||g) = \frac{d_i}{m} \int f(\upsilon,\sigma) \log \frac{f(\upsilon,\sigma)}{g(\upsilon,\sigma)} d(\upsilon,\sigma) \\ =& \frac{d_i}{m} \int f(\upsilon,\sigma) \log \frac{f(\upsilon,\sigma)}{\sum_{k=1}^{K} \beta_k \cdot g_k(\upsilon,\sigma)} d(\upsilon,\sigma),\end{aligned} \tag{2.20}$$

where $\frac{d_i}{m}$ represents the generality of subgroup i, which is a trade-off with exceptionality. Note that $g(\upsilon,\sigma)$ is a mixture of several distributions, with which it is difficult to compute the KL-divergence efficiently. In order to overcome this problem, we propose to compute the Goldberger approximation (Goldberger et al., 2003):

$$D_{\text{Goldberger}}(f||g) = \sum_{k=1}^{K} (D_{KL}(f||g_k) - \log \beta_k). \tag{2.21}$$

According to the properties of conjugate prior, the posterior distribution has the same form as the prior distribution. Thanks to properties of the $\mathcal{NG}$ function (Soch and Allefeld, 2016), we can compute the KL-divergence of two $\mathcal{NG}$ distributions as follows:

$$\begin{aligned}D_{KL_{\mathcal{NG}}}(f||g_k) =& \frac{1}{2}\kappa_{g_k}^2 \frac{\rho_f^2}{\psi_f^2}(\upsilon_{0g_k} - \upsilon_{0f})_+^2 \frac{1}{2}\frac{\kappa_{g_k}^2}{\kappa_f^2} - \log \frac{\kappa_{g_k}}{\kappa_f} - \frac{1}{2} \\ &+ \rho_{g_k} \log \frac{\psi_f}{\psi_{g_k}} - \log \frac{\Gamma(\rho_f)}{\Gamma(\rho_{g_k})} + (\rho_f - \rho_{g_k}) h(\rho_f) - (\psi_f - \psi_{g_k}) \frac{\rho_f}{\psi_f},\end{aligned} \tag{2.22}$$

where $h(x)$ is the digamma function. Combining this outcome with Equations (2.20) and (2.21), we compute the difference between the posterior distribution of time conditioned on one subgroup and the distribution of time in the whole dataset, denoted as φ_{t_i}. Similarly, we calculate φ_{l_i} and φ_{w_i}. Then we aggregate these three exceptionality indicators after normalizing to get the final exceptionality score:

$$\varphi_i = e^{\varphi_{l_i}^* + \varphi_{t_i}^* + \varphi_{w_i}^* - 3}. \tag{2.23}$$

2.6 Experiments

We evaluate the performance of our method on four real-world datasets from four cities on three continents: Twitter datasets from London, Tokyo, and New

Table 2.2: Datasets used in this chapter.

Dataset	# Tweets	# Users	Timeframe	# Attributes
London	169033	48232	April 2016	10
New York	210820	87510	April 2016	10
Tokyo	201643	49214	April 2016	10
Shenzhen	303161	100000	October 2016	8

York, and a Weibo dataset from Shenzhen. The details of datasets are shown in Table 2.2. The attributes of tweets contain: country, current living place, number of followers, number of following, listed, language, favourites, retweets, bio, date, source, gender, hour, latitude, longitude, and tweet text. We preprocess the tweets as follows:

1. converting the date into weekdays from 1 to 7;
2. extracting occupation from bio, such as student, driver, writer, editor, and so on;
3. removing stop words;
4. converting hours to float, from 1 to 24.

We use hour, latitude and longitude, and tweet text as the input values for temporal, spatial, and text information, respectively. All other attributes are used as the descriptors to generate subgroups. All the experiments are carried out on an Intel Core i7 2.60GHz laptop, 24GB RAM, Windows 10

To train BNPM by Algorithm 1, we must generate a set of input subgroups. To do so, we randomly sample 100,000 subgroups with replacement for which the coverages are ranging from 10 to 50 percent of the posts in the original dataset. For the spatial part, we calculate the mean coordinate and covariance from the data itself as the prior mean μ_0 and prior covariance W. The other hyper-parameters are initialized as follows: $\lambda = 1, \nu = 30$. For the temporal part, we calculate the prior mean of post time v_0 and initialize other hyper-parameters as follows: $\alpha = 0.1, \kappa = 0.1, \rho = 0.5, \psi = 0.1$. Through these settings and parametrizations, we train the BNPM model to capture the behavioral patterns in the global dataset; for instance the time distribution can now be estimated with Equation (2.19).

Having captured the global behavior, we can now mine for subgroups exhibiting exceptional behavior, by contrasting their behavior against the norm. We employ the beam search algorithm given in (Duivesteijn et al., 2016, Algorithm 1) for the subgroup search process. In the quality measure step, we calculate the exceptionality score of a subgroup by the method in Section 2.5.2. We set the beam width to 50 and the search depth to 2. This last parameter setting

Table 2.3: Exceptional subgroups in Shenzhen. We translate the original Chinese words into English, for your convenience. Descriptions: D_1: source == 'vivo', D_2: Gender == 'm' $\wedge$ source == 'other', D_3: source == 'vivo' $\wedge$ Gender != 'm', D_4: source == 'Mi' $\wedge$ Gender == 'm', D_5: Age >9 $\wedge$ Gender == 'm'. Higher $\varphi_{sd}(D)$ indicates more exceptionality. Higher $\frac{|D|}{|\Omega|}$ indicates more coverage of subgroup on the whole dataset.

D	$\varphi_{sd}(D)$	$\frac{\lvert D\rvert}{\lvert\Omega\rvert}$	High-Frequency Words
D_1	0.79	0.04	new song, come on, music, support, like, rank
D_2	0.64	0.04	Thailand, selfie, holiday, Weibo, tour, photography
D_3	0.62	0.03	new song, come on, music, support, like, rank
D_4	0.61	0.03	team, investment, customer, finance, refine, ability
D_5	0.51	0.04	stadium, sports, run, insist, seaside, struggle

Table 2.4: Exceptional subgroups in London. Descriptions: D_1: weekday:6-7 $\wedge$ Place == 'Hammersmith', D_2: Place == 'Camberwell', D_3: Place == 'Camden Town', D_4: Place == 'Hackney', D_5: Place == 'Kensington'

D	$\varphi_{sd}(D)$	$\frac{\lvert D\rvert}{\lvert\Omega\rvert}$	High-Frequency Words
D_1	0.95	0.03	London, Chelsea, Stamford, bridge, football, bar
D_2	0.90	0.07	stockmarket, trade, stock, intern, broker, forecast
D_3	0.88	0.07	street, kingcross, station, camdenlock, transport, driver
D_4	0.86	0.05	hackney, gym, class, image, orange, boss
D_5	0.85	0.04	history, restaurant, sweet, healthy, cover, Paddington

is relatively narrow; it ensures that we find subgroups expressed as a conjunction of at most two conditions on descriptive attributes. The reason to not mine to a greater search depth is philosophical rather than technical: computational complexity would allow us to mine deeper without prohibitive time cost, but when we allow our resulting subgroups to be defined in terms of a conjunction of more conditions on attributes, it becomes more and more opaque which of these conditions are actually relevant, and it becomes less clear what to do with the resulting information: mining deeper leads to subgroups which are no longer actionable.

London and Shenzhen In Table 2.3 and Table 2.4, we present the top 5 most exceptional subgroups found in Shenzhen and London, respectively. High frequency words in those subgroups are presented to show the main topics in the text of the tweets. We can see that the discovered subgroups restricted by spe-

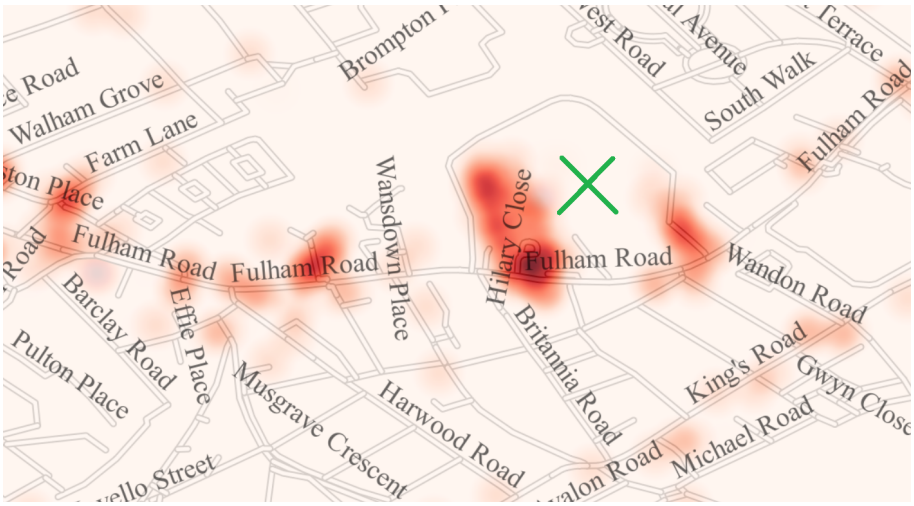


Figure 2.4: Spatial locations of tweets covered by description: "weekday:6-7 ∧ Place == Hammersmith London", plotted onto the map of London. The green cross highlights Stamford Bridge stadium.

cific descriptions show specific topical behavior, which can help us to further discover special events reflected by the group of social posts.

The top subgroup found in London encompasses the collective social posts described by "weekday:6-7 ∧ Place == Hammersmith London". The spatio-temporal behavior focuses on Saturday and Sunday in the borough of Hammersmith & Fulham in west London, a map of which is shown in Figure 2.4 with in red a heatmap of the spatial locations of the tweets. We visualize the texts of the posts by generating a word cloud shown in Figure 2.5, which shows that the main keywords of the tweets frequently contain *Chelsea*, *Stamford*, *Football*, *VS*, etcetera. It just so happens that on April 16, 2016, a Premier League football match between Chelsea and Manchester City was played at Stamford Bridge, which is the football stadium indicated by the green cross in Figure 2.4. Our model accurately captured this subgroup that has specific spatio-temporal behavior with specific word topics. This shows that our method can discover and identify meaningful exceptional collective behavior.

New York Figure 2.6 displays subgroups found in New York. Our method discovers a subgroup of people who live in Manhattan but do not speak English (D:Language != 'en' ∧ Place == Manhattan). From the word topics in those social posts, we can see that they are talking about the attractions and

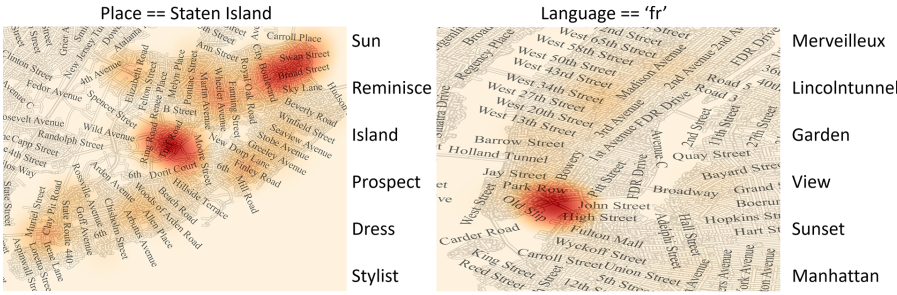
Figure 2.5: Word cloud generated from the texts of tweets covered by the subgroup plotted in Figure 2.4.

entertainments in Manhattan. In addition, we discover a subgroup of people discussing protest rallies in a suburb (D:Place == Yonkers), and a group of French speakers (Language == 'fr') sending tweets about a famous French backery, Aux Merveilleux de Fred. These findings show that characterizing groups of the dataset by the defined descriptive variables such as 'Language' and 'Place' contains sufficient information to discover subgroups with exceptional behavior in terms of spatial location, time, and texts.

Tokyo The full versatility of results that one could find with BNPM is on display in Figure 2.7, featuring the top subgroups found in Tokyo.

The top subgroup (D:Place == Chiyoda-ku) concentrates on the centrally-located special ward of Chiyoda. The heatmap shows that the people in this specific subgroup are mainly concentrated in three locations. The bottom-left location is the top attraction in Chiyoda ward: the imperial palace. The top-right location is Akihabara, nicknamed Akihabara Electric Town, which is a shopping district for video games, anime, manga, and computer goods; its function as a cultural center for all things electronic makes Akihabara a major touristic attraction in its own right. The bottom-right location is Tokyo station, which is far from a touristic attraction. Its relevance becomes clear when

(a)

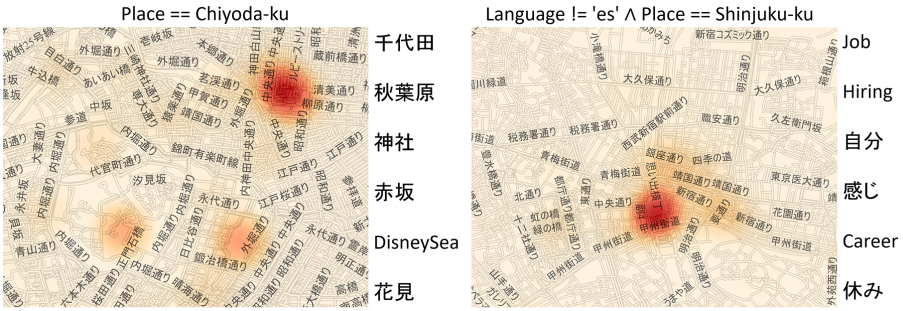


(b)

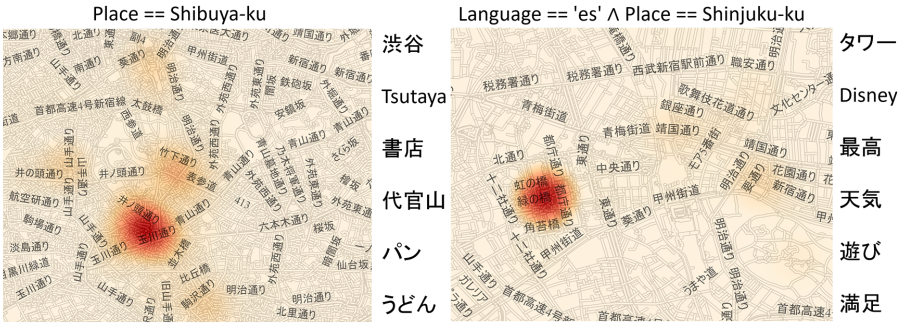
Figure 2.6: Most exceptional subgroups in New York; descriptions, maps, and high-frequency words.

looking at the tweet texts, which include references to DisneySea. This is yet another major touristic attraction of Tokyo, but it is located 15 kilometers away from Chiyoda ward. However, the easiest way for tourists to reach this destination is by taking a train on the Keiyo line, whose trains depart from Tokyo station. Hence, tourists who visit the imperial palace and Akihabara also express interest through tweets in visiting DisneySea, which is to be reached by a train departing from the ward in which the other two attractions lie. This finding shows that the combination of spatio-temporal behavior and word topics can benefit the discovery of such exceptional subgroups.

The second subgroup found in Tokyo (D:Language != 'es' ∧ Place == Shinjuku-ku) contrasts with subgroups discussed so far: these clearly are not tourists. Shinjuku is the major commercial and administrative center. Filtering out the people who tweet in Spanish (we will discuss this group later, in the fourth subgroup), we are left with a group of people discussing topics like job



(a)



(b)

Figure 2.7: Most exceptional subgroups in Tokyo; descriptions, maps, and high-frequency words.

hiring and career. Spatial locations of these people are strongly concentrated around Shinjuku train station (where big department stores, electronic stores, banks, and city hall are located), which makes sense for professionals.

The third subgroup (D:Place == Shibuya-ku) focuses on Shibuya ward, which is a major destination for fashion and nightlife. Arguably its most famous attraction is the Shibuya scramble crossing, a crosswalk at a busy intersection just outside of Shibuya station, where pedestrians in all directions (including diagonal) get the green light at the same time. The main spatial focus in this subgroup is located at that crossing. In the tweet texts we find references to Tsutaya, which is a book store located on a corner of that crossing. On the second floor of Tsutaya is a Starbucks coffee shop, whose numerous window seats overlook the scramble crossing.

In contrast with the second subgroup, the fourth subgroup found in Tokyo (D:Language == 'es' ∧ Place == Shinjuku-ku) concentrates on the same ward

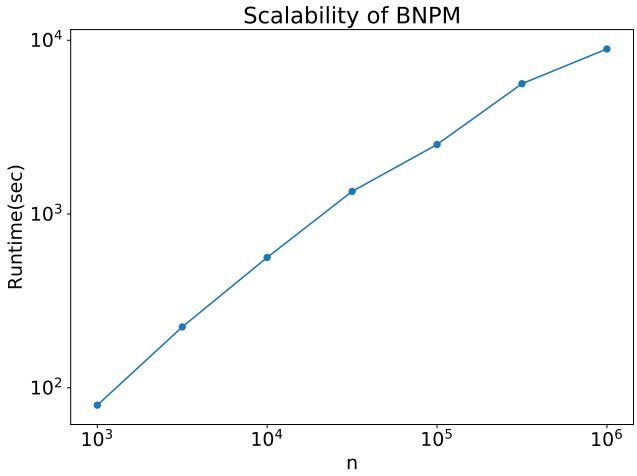


Figure 2.8: Runtime of **BNPM** vs. n.

(Shinjuku), but this time only on those people who tweet in Spanish. These are more likely to be tourists. The spatial location of these people is concentrated a few blocks to the west of Shinjuku station, where Tokyo Metropolitan Government Building is located. This building is famous for its observation deck, which provides a view over all of Tokyo and, if the weather is good, of Mount Fuji. This is the one place in Shinjuku which is of specific interest to tourists, and our BNPM model manages to separate out these from the professionals in the second subgroup. Notice also the interest expressed in the tweet texts of the fourth subgroup for Disney, which is absent from the tweets of the second subgroups.

Scalability In this chapter, we consider the scalability of our **BNPM** method in the aspect of model learning. The theoretical time complexity of Algorithm 1 is $\mathcal{O}(MAX \times n \times \bar{K})$. MAX represents the maximum number of loops we run random search for hyper-parameter optimization ($\bar{K}$ time) and collapsed Gibbs sampling ($n \times \bar{K}$ time). $\bar{K}$ represents the average number of latent components. n represents the number of input subgroups. Figure 2.8 shows the empirical relation between runtime behavior and n.

2.7 Conclusion

We propose a novel method for modeling multi-modal dependency and capture the uncertainty in multi-modal dependency. With this new method in EMM, it is possible to mine exceptional spatio-temporal behavior on collective social media. Behavior in this setting can be exceptional in three distinct ways: in terms of spatial locations, time, and texts. We develop a **B**ayesian **N**on-**P**arametric **M**odel (BNPM) to automatically identify spatio-temporal behavioral patterns on the subgroup level, explicitly modeling the three exceptional behavior types. Using a Chinese Restaurant Process, our model can cater for several distinct forms of global behavioral patterns, while also allowing for subgroup behavior that is exceptional with respect to all the kinds of global behavior. This behavioral dissimilarity can manifest itself in any subset of the three behavior types. The global distribution of the whole dataset can be summarized by the mixture of behavioral patterns with mixture coefficients in the components gathered by our model. We can also induce the distribution of a candidate subgroup by calculating its posterior distribution with BNPM, according to the behavioral data in that subgroup. The distance between the posterior distribution of the candidate subgroup and the global distribution indicates the exceptionality of that subgroup. This allows us to provide an effective evaluation method to measure the exceptionality of a behavioral pattern and to employ it in finding exceptional subgroups with collective social behavior. We develop an efficient learning algorithm based on collapsed Gibbs sampling to train the model.

We report results on datasets from various countries, continents, and cultures: BNPM finds exceptional subgroups in Shenzhen (cf. Table 2.3), London (cf. Table 2.4 and Figures 2.4 and 2.5), New York (cf. Figure 2.6), and Tokyo (cf. Figure 2.7). The results in London illustrate how BNPM can discovery unusual spatio-temporal tweeting behavior that coincides with a specific event: a Premier League football match of Chelsea F.C. (cf. Figures 2.4 and 2.5). But the capabilities of BNPM range far beyond event detection, as illustrated by the top subgroup found in Tokyo (cf. Figure 2.7, leftmost figure). Here, we discover a subgroup whose spatial behavior mostly revolves around three locations: two touristic attractions and a train station. The relevance of the train station becomes apparent when analyzing the tweet text behavior of the subgroup: the involved people frequently talk about a third touristic attraction 15 kilometers away, which is easiest reached by a train that departs from the discovered station. Hence, the exceptionality of this subgroup can only be properly appreciated by jointly analyzing the exceptionality of spatio-temporal

and tweet text behavior, which is precisely what BNPM is designed to do. Similarly, contrasting the second and fourth most exceptional subgroups found in Tokyo, we can distinguish the professionals from the tourists in Shinjuku ward by their exceptional joint spatial and tweet text behavior.

The four datasets analyzed in this chapter stem from four countries on three continents. Hence, we illustrate that BNPM is effective across various languages, religions, and cultures. In future work, it would be interesting to further investigate exactly how the vastly varying language patterns affect the proposed model.

3

Uncertainty in Dependency Modeling

"We cannot make the mystery go away by 'explaining' how it works. We will just tell you how it works. In telling you how it works we will have told you about the basic peculiarities of all quantum mechanics."

The Feynman Lectures on Physics,
Richard P. Feynman, 1963.

3.1 Introduction

The aim of dependency modeling is to capture how variables interact with each other. Exceptional Model Mining focuses on finding subgroups with unusual interactions, which can help us better understand the data and gain new knowledge from the exceptional patterns. In general, we model the dependencies between variables with functions (Caruccio et al., 2015); and define quality measures to compute the change and deviation between those functions in subgroups and the whole datasets. Challenges for this method are twofolds: on the one hand, datasets with complex structures usually have high dimensional interdependencies. Dimension reduction techniques are required to achieve low-dimensional representations of the interdependencies and to preserve original information as much as possible. On the other hand, observational datasets are usually imperfect, e.g. poor quality with missing information, or imbalanced data distribution. These challenges bring uncertainty to the dependency modeling and further influence the comparison of the dependencies in subgroups and the whole dataset. Properly overcoming these challenges can help us be aware of how much confidence do we have about the correctness and significance of the exceptional subgroups discovered by our algorithms. We develop systematic methods to capture the uncertainty in dependency modeling. Results are shown with two practical applications.

3.2 Practical Application in Fairness of Machine Learning

3.2.1 Motivation

There are increasing demands for machine learning on diverse real-world applications such as policing (Brennan et al., 2009), lending (Mahoney and Mohen, 2007) and credit scoring (Khandani et al., 2010). While recent advances in machine learning put many focuses on fairness of algorithmic decision making, topics about fairness of representation, especially fairness of network representation, are still underexplored. Fair decision making has become more and more important for machine learning research. Several notions have been defined for algorithmic fairness (Dwork et al., 2012, Hardt et al., 2016, Zafar et al., 2015). Among these methods, fairness is measured for individuals or pre-defined groups based on statistical quantities like false positive / negative rates or classification rates. Recently, more and more papers notice that the fairness of a decision making process is highly dependent on biases which already exist in the data collection process (Chen et al., 2018a). Network representation learning learns a function mapping nodes to low-dimensional vectors. Structural properties, e.g. communities and roles, are preserved in the latent embedding space. Fairness of representation learning receives a lot of attentions (Edwards and Storkey, 2015, Song et al., 2018, Madras et al., 2018). Among these methods, people are trying to learn similar representations for different groups, to ensure that the consequent decision making is independent of group attributes (Zhao and Gordon, 2019).

Despite the recent research focus on fair machine learning, the study of fair representation in networks still lacks exploration. Comparing with existing work, the challenges are two-fold: on the one hand, unlike statistical quantities of single decision variables, fairness of network representation requires to compare multi-degree interactions between nodes. We need to develop a new statistical measure to evaluate the differences between node representations. On the other hand, as pointed out by some research, when we only ensure fairness for some small amount of pre-defined subgroups, it might actually *increase* rather than decrease model discrimination (Kearns et al., 2017). In order to prevent this problem, we propose to investigate the fairness of network representation by generating subgroups with regard to any combinations of attributes. Computational cost would be very high due to the exponentially increasing amount of subgroups. We tackle this problem by employing Exceptional Model Mining (Duivesteijn et al., 2016), a framework of generating and

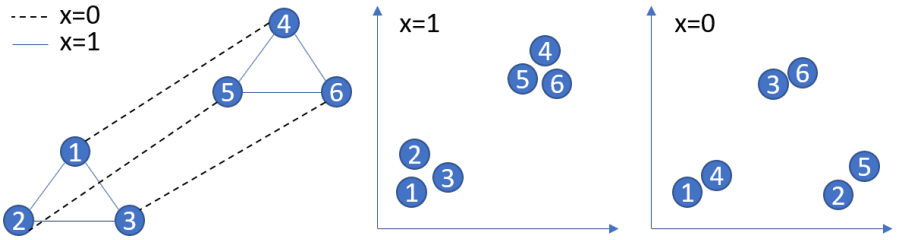


Figure 3.1: Toy example: dashed lines represent edges with attribute $x = 0$, solid lines represent edges with attribute $x = 1$. Obviously, the distributions of nodes in neighborhoods conditioned on different attributes ($P(N(V_o)|x = 1)$, $P(N(V_o)|x = 0)$) are different. This can lead to very different representation functions.

evaluating subgroups by heuristically exploring the attribute space.

Before discussing fairness of network representation, we firstly focus on *structural heterogeneity* in networks. Unknown heterogeneity across the data can lead a model to be very effective for some subpopulations and ineffective for some other subpopulations (Pearl, 2017). We argue in this chapter that the potential unfairness of network representation is associated with the structural heterogeneity in networks. In Figure 3.1, we demonstrate a toy example of structural heterogeneity and show how it can affect the network representation. As we can see, the network structure in subgroups '$x = 1$' and '$x = 0$' are very different from each other. A random walk based neighborhood function will generate different distributions of nodes in neighborhoods conditioned on different attributes. The classical network representation model could be biased. These biased representations might lead to unfairness of consequential decision making models. The study of fair machine learning should prevent the propagation of bias from the data to modeling results (Madras et al., 2019).

3.2.2 Contributions

In this section, we argue that latent structural heterogeneity in the observational data could bias the classical network representation model. The unknown heterogeneous distribution across subgroups raises new challenges for fairness in machine learning. Pre-defined groups with sensitive attributes cannot properly tackle the potential unfairness of network representation. We propose a method which can automatically discover subgroups which are unfairly treated by the network representation model. The fairness measure we propose can evaluate complex targets with multi-degree interactions. We analyze the latent struc-

tural heterogeneity across subgroups and discuss its effects on the fairness of network representations. Top-Q subgroups with highest measurement scores are reported to recover the fairness of a network representation model. In order to investigate whether the reported subgroups represent significant signals in the data, we conduct hypothesis testing against random noise. We conduct randomly controlled experiments on synthetic datasets and verify our methods on real-world datasets. Both quantitative and qualitative results show that our method is effective to recover the fairness of network representations. Our research draws insight on how structural heterogeneity across subgroups restricted by attributes would affect the fairness of network representation learning. The main contributions are:

- We study the problem of fairness in terms of the latent structural heterogeneity across subgroups in networks. As far as we know, this is the first work which considers structural heterogeneity to measure the fairness of network representation.
- We propose a new measurement, Mean Latent Similarity Discrepancy (MLSD) to quantify the differences between node representations. MLSD can calculate the statistical discrepancy between node representations which is sensitive to structural heterogeneity.
- We conduct hypothesis testing to verify the significance of fairness score, distinguishing structural discrepancy from randomized noise. We design a series of randomized experiments on synthetic and real-world datasets to evaluate our method qualitatively and quantitatively.

3.2.3 Related Work

Previous work on fair machine learning mainly focuses on the level of a group or individual. Pre-defined sensitive attributes are required, which is not applicable in many real-world applications (Kearns et al., 2017). Fairness on groups is normally measured by statistical parity, which requires positive / negative rate to be equal across groups with regard to sensitive variables (Hardt et al., 2016). Fairness on individuals requires similar individuals to be treated similarly by the models (Dwork et al., 2012). In contrast, fairness of network representation requires to compare more complex relations rather than a single decision variable. For this reason, we propose MLSD which focuses on measuring the statistical discrepancy between node representations.

Representation learning is specified to learn multiple degrees of similarities between units (Mikolov et al., 2013b) in large datasets. This technique is

Records	Descriptive Variables	Target Variables
r^1	$x_1^1, \ldots, x_k^1$	v_o^1, v_d^1
$\vdots$	$\vdots \quad \ddots \quad \vdots$	$\vdots$
r^n	$x_1^n, \ldots, x_k^n$	v_o^n, v_d^n

Table 3.1: A network dataset of N edges over a set of nodes $V = \{v_1, \ldots, v_m\}$ and attributes X= $\{x_1, \ldots, x_k\}$.

widely used to discover word similarities known as word embedding (Mikolov et al., 2013a) and node similarities known as graph embedding (Hamilton et al., 2017). Network representation learning enables us to learn low-dimensional vector representations for nodes from their neighborhood structures. There is a lot of work on learning vector representations of nodes in graphs (Perozzi et al., 2014, Grover and Leskovec, 2016) . Most existing work on fairness of representation focuses on adversely learning fair representations across groups and preserving highly predictive information for decision making (Zemel et al., 2013). Conversely, we focus on fairness of network representation, which requires definition of a new measurement with regard to the structural heterogeneity in networks. Our work can help people understand how structural heterogeneity is correlated with attributes and how unfairness of network representation exists by heuristically discovering subgroups.

Most of the existing model classes cannot handle structural properties in networks. Weighted relative accuracy was introduced to evaluate characteristics in subgraph (Bendimerad et al., 2016), first-order Markov chains have been introduced as a model class for sequential data (Lemmerich et al., 2016). However, structural properties, especially role structures (Jin et al., 2011) are not considered in those methods.

In order to compare the network representations which preserve the structural properties, we design the MLSD quality measure, based on the U-statistic (Korolyuk and Borovskich, 2013). MLSD calculates the mean discrepancy between latent similarities of node vectors, reflecting the statistical difference between network representations.

3.2.4 Methodology

Problem Setup We assume a dataset Ω: a set of M nodes $v \in V$ and a bag of N records $\mathbf{r} \in \Omega$ of the form $\mathbf{r} = (x_1, \ldots, x_k, v_o, v_d)$, where k is a positive integer and v_o, v_d refer to a directed edge from the origin v_o to the destination

v_d (cf. Table 3.1). We call $x_1, \ldots, x_k$ descriptive variables, and v_o, v_d target variables. The descriptive variables are taken from an unrestricted domain $\mathcal{A}$. Mathematically, we define descriptions as functions $D : \mathcal{A} \to \{0, 1\}$. A description D covers a record $\mathbf{r}^i$ if and only if $D(x_1^i, \ldots, x_k^i) = 1$. Subgroups and quality measure can be defined following definition 1.1.1 and 1.1.2.

We can model the network as $G_D = (V, E, X, D)$, where V represents set of M nodes, E set of N edges, X attributes attached on E, and D a description which is satisfied by X. We can define the neighborhood $N(v_o) \subset V$ as a set of nodes generated by a sampling strategy starting from node v_o. In this chapter, we consider local community structures, though our method can be easily extended to global role structure (Ribeiro et al., 2017). By defining the neighborhood function, we can formulate a distribution of nodes in neighborhoods conditioned on attributes $P(N(v_o)|D)$. If there is structural heterogeneity in networks, then we could have $P(N(v_o)|D) \neq P(N(v_o))$, and $P(N(v_o)|D_1) \neq P(N(v_o)|D_2)$ when $D_1 \neq D_2$. We would use this property to build the measurement for fairness of network representation.

By following the Skipgram model (Mikolov et al., 2013b), we can learn a function $\theta : V \to \mathbb{R}^l$, which maps each node $v \in V$ to a l-dimensional vector representation. We select θ_D to maximize the probability of visiting neighborhoods $N_D(v_o)$ for each node in network: $\theta_D = \operatorname{argmax}_{\theta_D} \prod_{v_o \in V} p(N_D(v_o)|\theta_D(v_o))$, where $\theta_D(v_o)$ can be represented as u_o. We can formulate the problem of fairness in network representation as an optimization problem of searching subgroups with highest quality scores:

Problem 3.2.1 Given a dataset $\Omega \sim P(X, V, E)$, a network representation model $\theta : V \to \mathbb{R}^l$, and a quality measure φ, our task is to find a sequence of Q descriptions $h = \{D_1, \ldots, D_Q\}$, such that $\forall D' \in \mathcal{D} \setminus h$, $\varphi(D') < \varphi(D)$, $\forall D \in h$.

Quality Measure: MLSD Node representations preserve the structural properties from the original networks. In order to measure the fairness across subgroups, we would like to evaluate the difference between node representations learned from that subgroup and learned from the whole dataset. To realize that, at first we need to elicit a latent similarity matrix Z_D, which indicates the similarities between each node and any other nodes:

$$Z_D^{ij} = \frac{d(u_i, u_j)}{\sum_{j \neq i}^{V} d(u_i, u_j)},$$

where $d(u_i, u_j)$ is a distance measure between node i and j in the latent embedding space, and $\sum_{j\neq i}^{V} d(u_i, u_j)$ is a normalizer that ensures $\sum_{j\neq i}^{V} Z_D^{ij} = 1$. Note that we do not consider self loop edges so we let $d(u_i, u_i) = 0$. Now we can compare the latent similarity matrix Z_D from candidate subgroup with Z_Ω to the whole data by using U-statistics (Korolyuk and Borovskich, 2013):

$$\varphi_u(D) = \frac{1}{m(m-1)} \sum_{i=0}^{m} \sum_{j\neq i}^{m} |Z_D^{ij} - Z_\Omega^{ij}|_1.$$

By virtue of variance, heterogeneous structures are likely to occur in small subsets of the dataset (Duivesteijn et al., 2016), which are not the results we want. To combat this problem, we incorporate the size of subgroups in the quality measure, by considering the entropy of the split between the records in subgroups and the rest of the records (Duivesteijn et al., 2010):

$$\varphi_{\text{ent}}(D) = -\frac{|D|}{n} \log_2 \left(\frac{|D|}{n}\right) - \frac{n-|D|}{n} \log_2 \left(\frac{n-|D|}{n}\right).$$

The final quality measure can be derived as:

$$\varphi_{\text{MLSD}}(D) = \sqrt{\varphi_{\text{ent}}(D)} \cdot \varphi_u(D).$$

By this quality measure, higher $\varphi_{\text{MLSD}}(D)$ indicates that the network representation is more unfair on that subgroup. By applying a search method guided by $\varphi_{\text{MLSD}}(D)$, we can derive the solution for problem 3.2.1.

Statistical Test In Problem 3.2.1, we report the top-Q subgroups with the highest scores calculated by quality measure. However, we do not know whether the scores are significant enough or just slightly different because of the random noise. To solve this problem, we assume that the reported vector of top-Q scores is a random draw from distribution P. We propose to independently run our method several times to generate a set of samples from P, denoted by $\mathbf{H} := \{h_1, \cdots, h_x\}$. On the other hand, we randomly shuffle the original data, by permuting the attribute vectors attached with edges in row (Batagelj and Brandes, 2005). This can break the dependencies between descriptive variables and targets, and build datasets where the descriptive variables are independent of network structures. After that, we apply our method on each of the shuffled datasets to generate false discoveries[1]. By doing this,

[1]Because now we already know the ground truth: the descriptive variables and network structures are independent.

we can generate a set of samples from the distribution of false discoveries (P_{DFD}) (Duivesteijn and Knobbe, 2011), denoted by $\tilde{\mathbf{H}} := \{\tilde{h_1}, \cdots, \tilde{h_y}\}$. Now we can build the null hypothesis by assuming that $\mathbf{H}$ and $\tilde{\mathbf{H}}$ are from the same distribution:

Hypothesis 3.2.1 P and P_{DFD} are the same distribution.

If the null hypothesis is rejected, then we can be confident that the top-Q subgroups reported by our method are statistically significant. We can define the problem as:

Problem 3.2.2 Let h and $\tilde{h}$ be random variables defined on a topological space $\mathcal{H}$, with distribution P and P_{DFD}. $\mathbf{H} := \{h_1, \cdots, h_x\}$ and $\tilde{\mathbf{H}} := \{\tilde{h_1}, \cdots, \tilde{h_y}\}$ are defined as independently and identically distributed samples from P and P_{DFD} respectively. The problem is to establish a statistical test and conduct hypothesis testing to decide whether $P = P_{DFD}$.

The main challenge for Problem 3.2.2 is that h and $\tilde{h}$ are multivariate (Q-length) and we do not have any prior knowledge about distribution P and P_{DFD}. Hence, classic Student's t-test and Hotelling's T^2-test are not appropriate. Inspired by (Gretton et al., 2012), we use an integral probability metric (Müller, 1997) based on distances between Hilbert space mean embeddings of probability distributions, termed as maximum mean discrepancy (MMD). Let $\mathcal{F}$ be a family of functions $f : \mathcal{H} \to \mathbb{R}$, we have:

$$MMD[\mathcal{F}, P, P_{DFD}] := \sup_{f \in \mathcal{F}} (\mathbb{E}_P[f(h)] - \mathbb{E}_{P_{DFD}}[f(\tilde{h})]),$$

where h and $\tilde{h}$, P and P_{DFD} follow Problem 3.2.2. Empirically, we can derive the unbiased estimate of the squared MMD in terms of kernel functions ψ as:

$$MMD_u^2[\mathcal{F}, \mathbf{H}, \tilde{\mathbf{H}}] = \frac{1}{x(x-1)} \sum_{i=1}^{x} \sum_{j \neq i}^{x} \psi(h_i, h_j) +$$
$$\frac{1}{y(y-1)} \sum_{i=1}^{y} \sum_{j \neq i}^{y} \psi(\tilde{h}_i, \tilde{h}_j) - \frac{2}{xy} \sum_{i=1}^{x} \sum_{j=1}^{y} \psi(h_i, \tilde{h}_j),$$

which is a sum of two U-statistics and a sample average. Following (Anderson et al., 1994), we would like to use asymptotic distribution of MMD_u^2 under null hypothesis for the hypothesis testing, by assuming that P and P_{DFD} are identical. Hence if we generate two new data samples from the aggregated data

samples after random shuffle, the MMD_u^2 should not change. We can construct null distribution by re-shuffling the aggregated data samples and re-computing the MMD_u^2 a lot of times. Given a significance level α, if MMD_u^2 is so large as to be outside the $1 - \alpha$ quantile of the null distribution, we can reject the null hypothesis, otherwise we accept it.

3.2.5 Experiments

In this section, we design synthetic and real-world experiments to validate our methodology against the following questions:

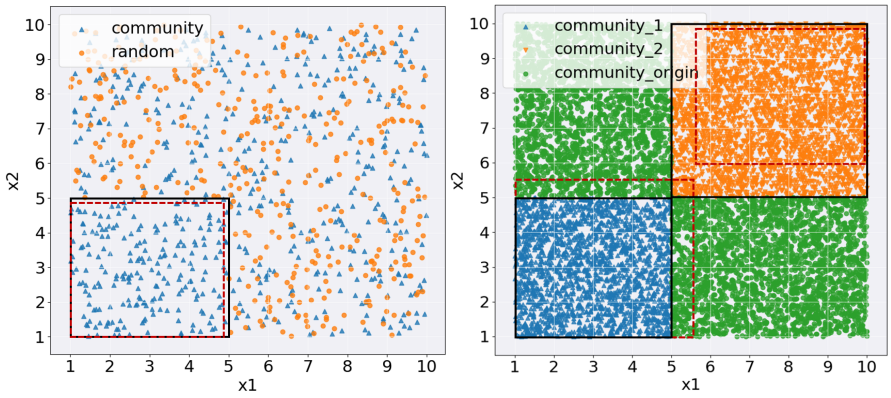
QS1 When there exists a latent structural heterogeneity, will the classical network representation model like node2vec perform fairly across different subgroups?

QS2 Can our method effectively measure fairness of network representation considering structural heterogeneity in subgroups?

QS3 Are the fairness measurement scores reported by our method significant enough compared to the random noises?

The most difficult problem for evaluating our methods is the lack of ground truth. For an observational dataset, we do not know whether there is structural heterogeneity and consequently we cannot know whether we can correctly measure the fairness. To overcome this, we design experiments with synthetic data generated by controlling the dependencies between descriptive variables and the network structures. By doing this, the experiments can evaluate the performance of our method by comparing them with the ground truth. For real-world datasets, we will never know the ground truth, but the statistical test can help us to evaluate the methods against the random baselines. Qualitative and visual analysis can be used to show the effectiveness of the discoveries.

Synthetic datasets with ground truth As synthetic datasets, we employ modified versions of the two datasets from (Girvan and Newman, 2002). The two datasets are called *Karate* and *Football*. We keep the original nodes and community label and drop all the connections. The generating process of the synthetic datasets is governed by following parameters: the number of records N, the number of descriptive variables K, the set of nodes V, and the set of ground truth labels Y indicating communities. We propose a randomized



(a) KarateX4n1k. **(b)** FootballX4n10k.

Figure 3.2: Randomized synthetic datasets with ground truth. Rectangles with solid lines denote ground truth subgroups. Rectangles with dash lines denote the subgroups reported by our method.

technique to model the dependencies between target variables v_d, v_o and descriptive variables $x_1, \ldots, x_k$. Two kinds of heterogeneous structures are generated: one is a community structure in subgroups against uniform distribution of edges in global, another is a core-periphery structure (Borgatti and Everett, 2000). We visualize two examples 'KarateX4n10k' (K=4, N=10,000, $|V|$=34) and 'FootballX4n10k' (K=4, N=10,000, $|V|$=115) in Figure 3.2. In Figure 3.2a, triangles represent the edges inside communities and dots represent uniform sampled edges between any pair of nodes. We can see that blue triangles are distributed uniformly except in the black rectangle. In the ground truth subgroup, the edges only exist in the local community. In Figure 3.2b, we synthesize a simple core-periphery structure. This is one of the simplest global role structures which consists of dense and cohesive core nodes as well as sparse and unconnected periphery nodes.

Real-world datasets As real-world datasets, two kinds of data are used for the experiments: (1) the original edge connections; and (2) extra data about the contextual information. We collect the original edge connections including 'New York Taxi' (`http://www.nyc.gov/html/tlc/`) (K=33, N=1,013,845, $|V|$=265) and 'Sharing Bike' (`https://datasf.org/opendata/`) (K=27, N=983,000, $|V|$=70), as well as the contextual information, e.g. weather records (`https://www.ncdc.noaa.gov/`) and taxi

KarateX4n10k

D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$
$x_1 \leq 4.86 \wedge x_2 \leq 4.86$	.0225	.188
$x_1 \leq 3.57 \wedge x_2 \leq 4.86$	.0224	.188
$x_1 \leq 4.86 \wedge x_2 \leq 3.57$	.0201	.128
$x_1 \leq 3.57 \wedge x_2 \leq 3.57$	.0196	.123
$x_1 \leq 6.14 \wedge x_2 \leq 4.86$	.0076	.249

FootballX4n10k

D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$
$x_2 \geq 6.14 \wedge x_1 \geq 4.86$	.0047	0.244
$x_2 \geq 6.14 \wedge x_1 \geq 6.14$	.0015	0.182
$x_2 \leq 4.86 \wedge x_1 \leq 4.86$	.0015	0.182
$x_1 \leq 4.86 \wedge x_2 \leq 3.57$	.0014	0.124
$x_1 \leq 3.57 \wedge x_2 \leq 4.86$	.0014	0.124

Table 3.2: Top-5 subgroups discovered on KarateX4n10k. The higher $\varphi_{\text{MLSD}}(D)$, the more unfair. $\frac{|D|}{N}$ indicates the coverage of subgroups.

information. By choosing these two datasets, we would like to show how network representation model can be biased by attributes like weather conditions. Consequently the downstream tasks (e.g. transportation prediction on specific weather conditions) could also be biased using the representations learned from the whole data. From these experiments we show that study for fairness of network representation has broad application fields.

Implementation details For the implementation of node representation learning, we build the algorithm based on Node2vec (Grover and Leskovec, 2016). For each candidate subgroup, we construct the graph with edges covered by that subgroup and use a random walk algorithm considering the aggregated edge weights to generate the training labels. After getting the node representations, we compare them with node representations learned from the whole data. To explore the attribute space with exponential amounts of subgroups, we use beam search guided by the quality score heuristically. The beam search algorithm is built based on (Duivesteijn et al., 2016, Algorithm 1). We set the beam width to 5 and depth to 2. All the experiments are conducted on Linux computing clusters with CPU: 2x Intel Xeon @ 2.1GHz and RAM: 1024GB.

KarateX4n10k			FootballX4n10k		
Q	TPR	PPV	Q	TPR	PPV
5	.61	.94	5	.69	1.0
10	.40	.86	10	.53	.96
25	.36	.71	25	.44	.52
35	.36	.65	35	.28	.51
50	.33	.50	50	.28	.50

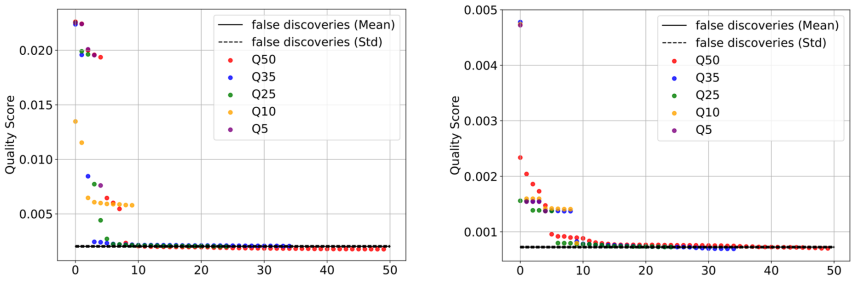
Table 3.3: Experimental results on synthetic datasets. The higher TRP and PPV the better.

Experiments on Synthetic Data To validate our method against QS1 and QS2, we conduct experiments on the two synthetic datasets with different settings mainly by varying parameter Q, which indicates how many subgroups we are going to report. The top-5 subgroups are reported in Table 3.2. As shown in Figure 3.2, our algorithm can discover the pre-imposed structures with good accuracy.

The subgroups we found cannot always be precisely the ground truth. The rectangles with black solid lines and the rectangles with red dot lines are slightly mismatching (cf. Figure 3.2). There might be two reasons for that. On the one hand, we employ a 8-bin equal-width binning strategy to partition the space of descriptive variables denoted by continuous numerical values. On the other hand, we prune the result set based on overlapping coverage to reduce redundant discoveries. Hence, we plan to evaluate more about the predictive ability of our method. According to the known label of each edge, we can calculate averaged number of edges covered by discovered subgroups to build the confusion matrix. We choose true positive rate (TPR) and positive predictive value (PPV) as the evaluation indicators.

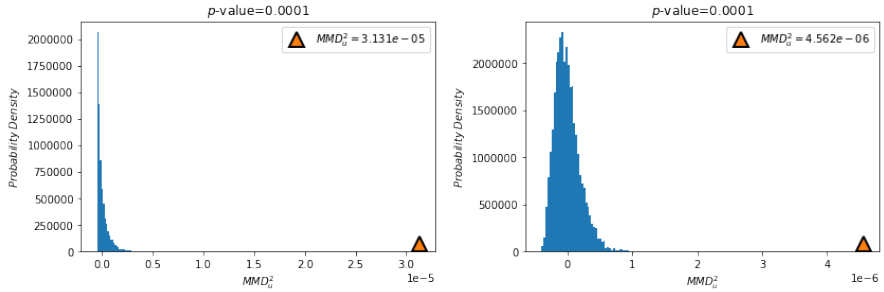
Table 3.3 displays the results; larger TPR and PPV indicate better results. We can see that for the same dataset, with the increasing of Q, MMD_u^2, TPR and PPV decrease. One reason for this phenomenon is that the forced diversity of discovered subgroups works against identification of the single ground truth subgroup. Another reason is that larger Q allows for subgroups with lower qualities, so that some records without label of ground truth are discovered by our method. We also notice that the PPV of finding subgroups by our method are always larger than 50%, which shows that our method can reliably retrieve ground-truth subgroups.

In order to validate our method against QS3, we run our algorithm on the



(a) KarateX4n10k. **(b)** FootballX4n10k.

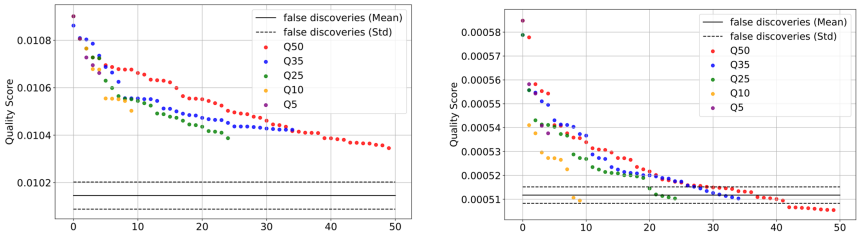
Figure 3.3: Comparisons of quality score distributions.



(a) KarateX4n10k. Q=10. **(b)** FootballX4n10k. Q=10.

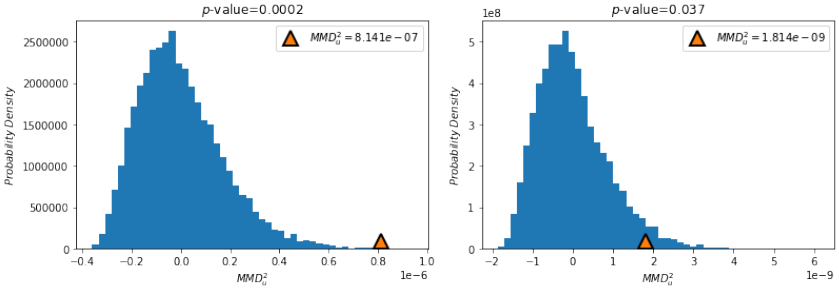
Figure 3.4: Visualization of null distribution and MMD_u^2 on KarateX4n10k and FootballX4n10k datasets.

randomly shuffled datasets for 100 times to generate negative samples. In Figure 3.3, we plot the quality scores in different experiments with Q ranging from 5 to 50, as well as the quality scores from negative samples. We can see that there is a large gap between quality scores of reported subgroups and the false discoveries. One reason is that with synthetic algorithm, we impose very different structural properties. Also we noticed that there are many low ranked subgroups dropping into the region of false discoveries. The reason is that the number of pre-imposed discriminated subgroups are less than the Q. Then we conduct the hypothesis testing to investigate whether the differences between our discoveries and the false discoveries are significant enough. In Figure 3.4, we visualize the null distribution and report p-value with $Q = 10$ on KarateX4n10k and FootballX4n10k. As we can see intuitively, the MMD_u^2



(a) Quality distribution bike. **(b)** Quality distribution taxi.

Figure 3.5: Quality score comparisons on dataset Sharing Bike and New York Taxi.



(a) Bike. Q=10. **(b)** Taxi. Q=10.

Figure 3.6: Visualization of null distribution and MMD_u^2 on bike and taxi datasets.

is far from null distribution. We can be confident that our method can beat false discoveries generated from random baselines. We also noticed that based on the p-values we can reject the null hypothesis at 1% significance level.

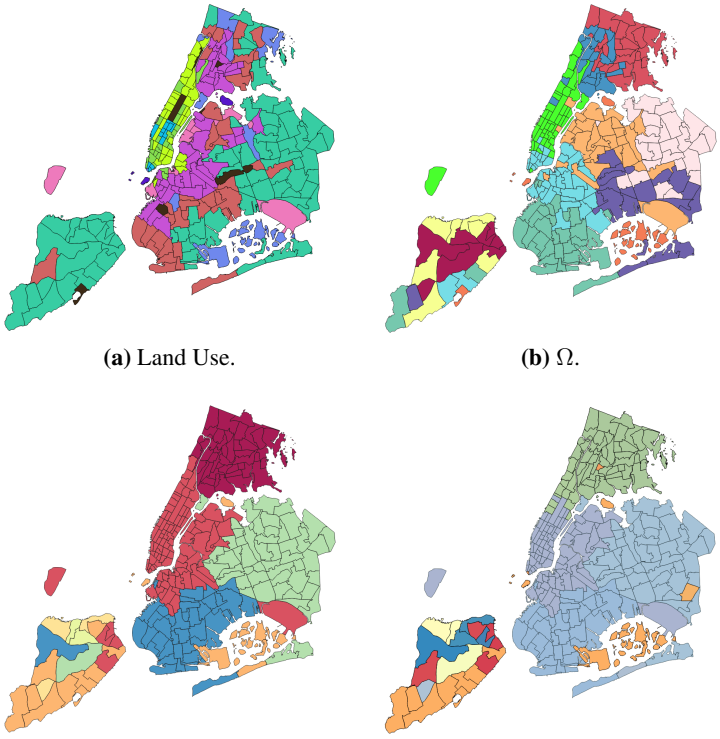
Experiments on Real-world Datasets Similar experiments are conducted on the real-world datasets, except calculating TRP and PPV due to the reason that we do not know the ground truth. In Figure 3.5, we plot the quality scores of discovered subgroups in different experimental settings with Q ranging from 5 to 50. We can see that in the real-world datasets, the quality decreases more smoothly than in the synthetic. One reason might be that in the real-world datasets, there are many kinds of combinations between structural properties and descriptive variables. Another reason might be that the attribute space and number of edges are much larger than the synthetic datasets so that the performance of network representation models are more diverse. As we

Dataset	D	$\varphi_{\text{MLSD}}(D)$	$\frac{\lvert D\rvert}{N}$
Sharing Bike	MaxHumidity $<=$ 74.0 $\wedge$ ZipCode ! = '10010'	.01090	.194
	MinTemperatureF $>$ 50.0 $\wedge$ MaxTemperatureF $>$ 70.0'	.01081	.232
	MaxHumidity $<=$ 74.0 $\wedge$ ZipCode ! = '7050'	.01073	.194
	MaxHumidity $<=$ 74.0 $\wedge$ ZipCode ! = '77450'	.01069	.194
	MaxHumidity $<=$ 74.0 $\wedge$ ZipCode ! = '19119'	.01066	.194
New York Taxi	month $>$ 7.0 $\wedge$ PaymentType $<=$ 1.0	5.85e-4	.211
	TMIN $>$ 61.0 $\wedge$ PickupHour $<=$ 14:00	5.58e-4	.126
	month $>$ 7.0 $\wedge$ AWND $<=$5.24	5.54e-4	.272
	month $>$ 7.0 $\wedge$ TMIN $>$ 42.0	5.41e-4	.279
	month $>$ 7.0 $\wedge$ TMAX $>$ 54.0	5.38e-4	.300

Table 3.4: Experiments on real-world datasets. Higher $\varphi_{\text{MLSD}}(D)$ means more unfair.

can see in Figure 3.6, the MMD_u^2 and p-values give us confidence to believe that there are significant differences between the subgroups reported by our method and the false discoveries. In Table 3.4, we report the top-5 subgroups in both datasets. We can see from the descriptions that the weather conditions and urban regions are highly related with the heterogeneous structures. This indicates that the decision models might be more vulnerable and discriminated under such conditions.

Empirical Clustering Analysis To further explore these results, we conduct clustering on taxi zones in New York using k-means algorithm with the learned representations from taxi transitions. We use the discovered subgroups above and the whole dataset as the input to train representations for each taxi zone. On the one hand, we would like to see how these clusters are different between reported subgroups and the whole dataset. On the other hand, we would like to see how the representations of taxi zones are changing with the changing of descriptive variables.

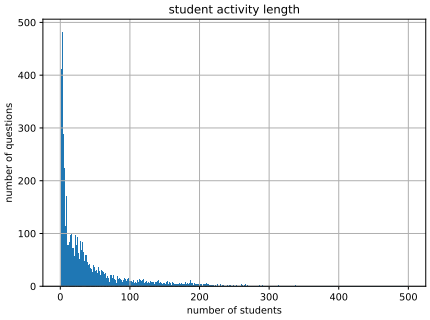


(a) Land Use. **(b)** Ω.

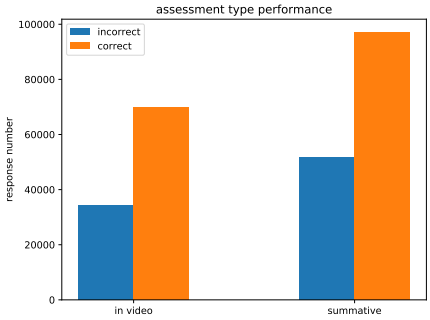
(c) 'SNWD$>$0.0$\wedge$AWND$\leq$ 7.86'. **(d)** Passenger$>$ 5.

Figure 3.7: Taxi zone clusters with representations.

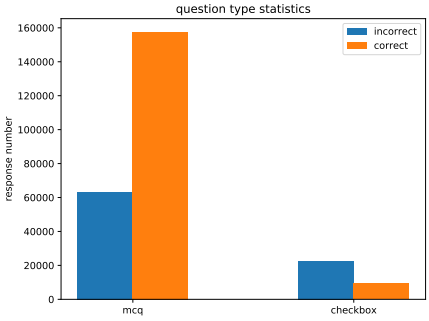
To conduct this comparison, we employ the land use data in New York (`https://zola.planning.nyc.gov/`) as a reference of the ground truth. The assumption is that taxi zones with similar land use types are similar to each other. Based on this assumption, we count the land use types in each taxi zone, and compute the distribution of land use types as the representation of each taxi zone. We visualize these clustering results in Figure 3.7. By comparing those clusters in Figure 3.7a with the clusters learned on the whole dataset (cf. Figure 3.7b), we found the similarities between taxi zones can be preserved relatively well. In Figure 3.7c, John F. Kennedy International Airport shows different role with nearby zones, while it shows the same role with the Manhattan area. In Figure 3.7d, we can see that for 'passenger $>$ 5', many zones that are distinguished in previous subgroups become more similar. These results empirically show the structural heterogeneity in different subgroups. For fair decision making, a network representation model should



(a) Histogram of the number of assessment questions in which students participate.



(b) Performance per assessment type.



(c) Performance per question type. 'mcq' represents multiple checkbox questions.

Figure 3.8: Heterogeneity and inconsistency of student behavior.

tackle this heterogeneity to learn fair as well as informative representations.

3.3 Practical Application in Educational Data Mining

Behavioral records collected through course assessments, peer assignments, and programming assignments in Massive Open Online Courses (MOOCs) provide multiple views about a student's study style. Study behavior is correlated with whether or not the student can get a certificate or drop out from a course. It is of predominant importance to identify the particular behavioral patterns and establish an accurate predictive model for the learning results, so that tutors can give well-focused assistance and guidance on specific students. However, the behavioral records of individuals are usually very sparse;

behavioral records between individuals are inconsistent in time and skewed in contents. These remain big challenges for the state-of-the-art methods. In this section, we engage the concept of *subgroup* as a trade-off to overcome the sparsity of individual behavioral records and inconsistency between individuals. We employ the EMM framework to discover exceptional student behavior. Various model classes of EMM are applied on dropout rate analysis, correlation analysis between length of learning behavior sequence and course grades, and passing state prediction analysis. Qualitative and quantitative experimental results on real MOOCs datasets show that our method can discover significantly interesting learning behavioral patterns of students.

3.3.1 Motivation

Massive Open Online Courses (MOOCs) make it possible for educators to analyze learning behavior of students in multiple views. In contrast to traditional classes, which only have limited learning behavioral records, MOOC platforms such as Coursera, edX and Udacity provide huge amounts of learning behavioral records. These platforms collect very detailed course information and students' learning behavior such as course assessments, peer assignments, programming assignments, forum discussions and feedback (Seaton et al., 2014), which can reflect the knowledge and skill achievements and the study performance of students. Modeling students' learning behavior and trying to discover interesting behavioral patterns are non-trivial tasks. Most recent research is focused on how to predict the learning results based on the learning behavior model. It can help the tutors to design the courses and give specific guidance and assistance to specific students. However, due to the complexity of the behavioral records, there remains several challenges to overcome:

Individual sparsity. Even when many students are enrolled in a course, the duration of their involvement varies substantially. Figure 3.8a displays a histogram of assessment question frequencies, which shows an obvious Power-Law distribution (Barabási and Albert, 1999). Only a few students participate in hundreds of assessment questions. Most of the students have activity length less than 20 records, which is very sparse. This makes evolutionary activity sequence based user modeling methods (Qiu et al., 2016, 2013) ineffective.

Activity inconsistency. Beyond the distribution in activity length of assessment questions, students' learning behavior in forum discussion, click stream

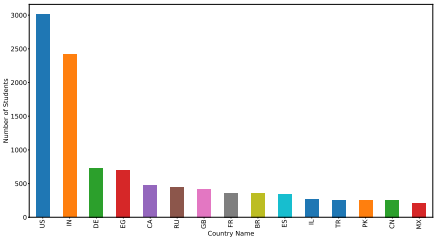
and peer review are also shown to follow a Power-Law distribution. In Table 3.7, we can see that among the 18 courses on Coursera, enrolled students, grades and students who passed the course are highly diverse. This inconsistency makes the data very imbalanced, which results in difficulties for Matrix factorization based modeling methods (Zhao et al., 2015). These methods might merge infrequent behaviors with common behaviors.

Content heterogeneity. Behavior diversity is not only shown in activity length and course status, but also shown in informative contents. There are 7 types of assessments and 12 types of questions in the courses, such as video, summative, checkbox and multiple checkbox. Proportions of these assessments and questions are skewed in different courses. On the other hand, students also have varying participation records on these contents. In Figure 3.9, it is shown that distributions of students are obviously different in specific demographic categories. It is a big challenge for modeling methods to handle these heterogeneous contents for tasks like dropout prediction or passing state prediction.

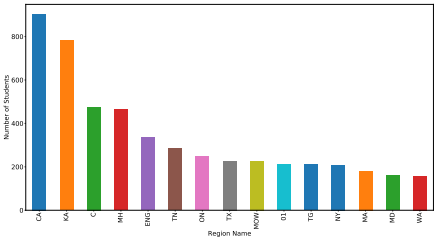
3.3.2 Contributions

To overcome these challenges, we propose to employ EMM for exceptional learning behavior analysis. Instead of looking for anomalies or outliers of individuals, we look for exceptional behavior on the subgroup level, which can provide interpretable descriptions such as 'Students: Country = US, Region = Manhattan, Join dates $>$ 365 (days)' having exceptional learning behaviors that are predominantly different from those in the whole dataset. We employ EMM to discover interesting learning behavioral patterns in subgroups. We establish various model classes for specific learning behaviors, such as discovering correlation between length of behavior sequence and course grades, finding out subgroups with exceptional dropout ratio, and looking for specific subsets where the classifier does not perform well. Experimental results on a real dataset illustrate the type of meaningful learning behavioral patterns EMM can discover in MOOCs. This can help us build an improved behavior model in the future research. In summary, our main contributions are:

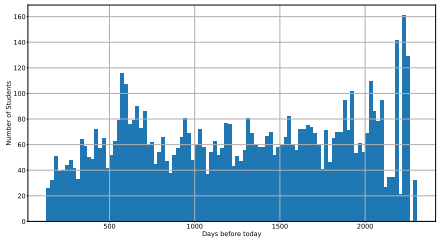
- We employ EMM to learning behavior analysis in MOOCs, which can help us to overcome the sparsity, inconsistency and heterogeneity in the behavioral records.
- We employ several EMM model classes for different tasks to discover



(a) Number of students by country.



(b) Number of students by regions.



(c) Number of students by join date.

Figure 3.9: Student distributions across various demographic categories.

exceptional learning behaviors on the subgroup level. Our results show very interesting learning behavioral patterns, which can help the tutors conduct specific guidance and assistance to the students.

3.3.3 Related Work

Learning behavior modeling for students in MOOCs is generally aimed at predictive analytics such as dropout prediction, passing state prediction, and grades prediction. For instance, latent factors and state machines are employed to model the hidden study state of students for a predictive task (Ramesh et al., 2014, Qiu et al., 2016, Wang and Chen, 2016). Khajah et al. (Khajah et al., 2014) integrate Latent factor and knowledge tracing with a hierarchical Bayesian model, which can consider the study skill for prediction tasks. Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have been used to model study trajectories for the learning results prediction (Piech et al., 2015, Wang et al., 2017). Most of these existing methods focus on modeling individual behavior but do not consider the sparsity, inconsistency and heterogeneity of learning behavior data. Our methods focus on discovering exceptional learning behaviors on the subgroup level, which provide inter-

pretable information about where the predictive model does not perform well. This allows us to establish an improved model for prediction tasks for both normal and exceptional behavioral patterns.

3.3.4 Exceptional Learning Behavior Analysis

Our dataset originates from the learners involved in the EIT Digital MOOCs at Coursera. EIT Digital, as part of the European Institute for Innovation and Technology, aims to drive Europe's digital transformation, also for education. The EIT Digital academy is focused on mobility and entrepreneurship and is at the forefront of integrating education, research, and business. The MOOCs in the online programme, have been developed by the partner universities involved in the EIT Digital Master School in Embedded Systems, in a best of breeds approach.

Together, the MOOCs form the EIT Digital online programme "Internet of Things through Embedded Systems". The online programme aims to build the reputation of EIT Digital, the partner universities, and the involved teachers. It also helps to renew pedagogy through scalable education technologies and data driven education. Learning analytics are at the core of this feedback mechanism. The online programme is comparable to an edX's micromaster and similarly offers an online equivalent of a 25 ECTS first semester; the online programme offers learners to study at their own pace, any time, any place. Moreover, they first can have a try before they commit themselves to the whole master programme. Once selected and admitted on campus, the learners can finish the double degree master programme of EIT Digital Master School in Embedded Systems.

Figure 3.9 displays the distributions of students across various demographic categories. In order to catch the inherent imbalance, we use demographic columns as the left hand attributes, to formulate subgroup descriptions. In the data preprocessing process, we convert the join dates, which represents how long a student has registered in Coursera, from the format of 'Datetime' to the integer days. The following three sections illustrate what kind of discoveries can be made by wielding various tools from the EMM toolbox.

Exceptional Dropout Rate Analysis In this section, our task is to find out the subgroups which have significantly different dropout rate compared with the whole dataset. For the purposes of this section, we define a dropout student to be a student who has participated in at least one assessment question, but has

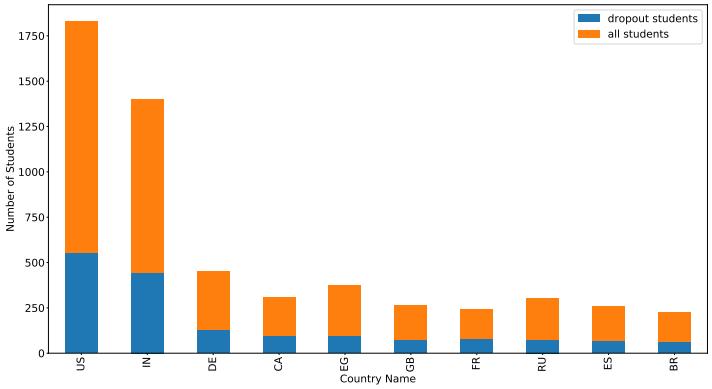


Figure 3.10: Dropout ratio of students by country.

not obtained an overall course grade. In Figure 3.10, we present the countries of students with most common frequencies, as well as the dropout rate of students in those countries. We can see that both the frequencies and dropout rate in those countries varies a lot. The high dropout rate is usually seen as a defect of MOOCs. If we were to discover what kinds of students have exceptional dropout rates, then that would allow us to direct specific guidance to those students that most require it. Traditional partition and clustering methods are not qualified for this task, because they cannot provide interpretable results about the subsets of students and quantitative information about how different the subsets of students are from the whole dataset. To address this problem, we propose to engage subgroups as a partition for the whole dataset, and look for subgroups that have most exceptional dropout rate comparing with the whole dataset. To this end, we employ *Weighted Relative Accuracy* (WRAcc) (van Leeuwen and Knobbe, 2011):

$$\varphi_{\text{WRAcc}} = \frac{|G_D|}{N}\left(\frac{S_D}{|G_D|} - \frac{S_\Omega}{N}\right)$$

Here, $|G_D|$ represents the number of records covered by subgroup description D, S_D represents the number of dropout students in subgroup G_D, S_Ω represents the total number of dropout students in the whole dataset, and N represents the number of students who join this course and participated in at least one assessment question.

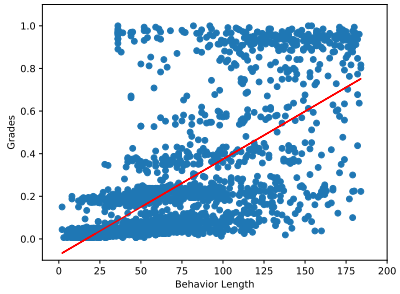
The beam search algorithm as described in (Duivesteijn et al., 2016, Algorithm 1) is parameterized with beam width 20 and search depth 4. The overall dropout rate is 0.4286. In Table 3.5, we presents the top-5 subgroups with

Table 3.5: Exceptional dropout rate in subgroups. Results show subgroups with highly exceptional dropout rate. The overall dropout rate is 0.4286.

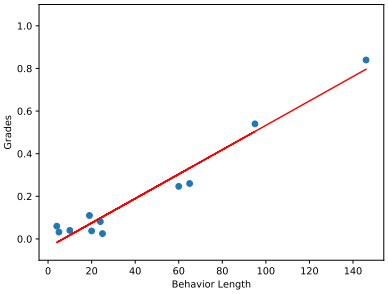
D	φ_{WRAcc}	dropout	$\lvert G_D \rvert$
Country = OM, Was Group Sponsored ¬ True, Was Finaid Grant ¬ True	0.0338	0.0	42
Region = MOW, Gender ¬ male, Join Date $\leq$ 1011, Join Date $>$ 389	0.0336	0.0	57
Country = KR, Gender ¬ female, Profile language ¬ ko	0.0330	0.7812	32
Country = KR, Educational status ¬ MASTER, Gender ¬ female, Was Group Sponsored ¬ True	0.0313	0.7742	34
Country = KR, Was Group Sponsored ¬ True	0.0304	0.7222	36

most exceptional dropout rate. The subgroup with description "D: Region = MOW, Gender ¬ male, Join Date between 389 and 1011" has a dropout rate of zero: all students in that subgroup complete the course. On the other hand, the subgroup with description "D: Country = KR, Gender ¬ female and Profile language ¬ ko", has an elevated dropout rate of 0.7812: most of these students drop out. Based on these results, we can conclude that Korean males who have set their profile language to something other than Korean, are in need of more attention. This may be a group of students who are foreigners in Korea, or Koreans who are studying in a language which is non-native to them. By identifying such at-risk groups, educators can more effectively channel their remedial activities.

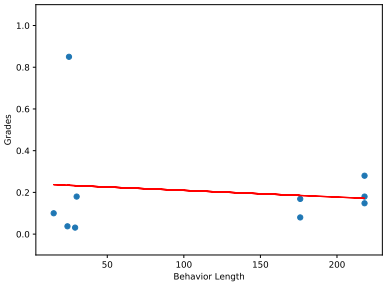
Exceptional Correlation Analysis In general opinion, the more active a student is, the higher grades she is promising to get. Is this always the case? To answer this question, we look into the relation between the activity length (denoted by q) of students and the overall grades (denoted by g) in a course. We engage the correlation model class for EMM to realize this task. In this model class, we can estimate the correlation coefficient by calculating the sample



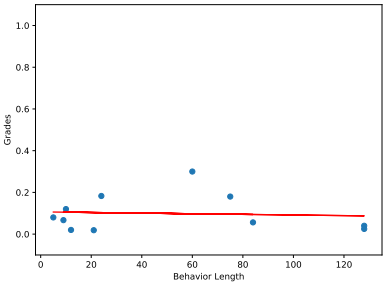
(a) The whole dataset. $\rho = 0.7406$



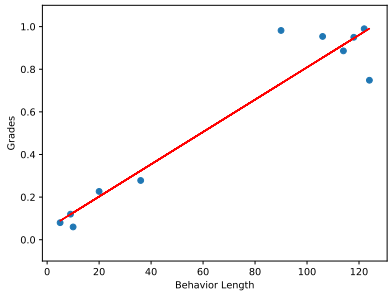
(b) Country = LT, Join Date $>$ 701, Browser language $\neg$ et-EE. $\rho = 0.9782$



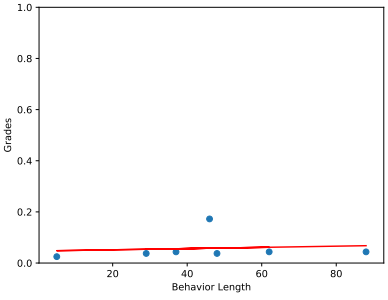
(c) Region = 6. ρ = -0.1272



(d) Region = QUE. ρ = -0.0788



(e) Country = NP. $\rho = 0.9630$



(f) Browser language = es-MX. $\rho = 0.1203$

Figure 3.11: Exceptional correlations in subgroups.

Table 3.6: Exceptional correlation analysis between length of behavior sequence and course grades. The overall correlation coefficient ρ is 0.7406.

D	φ_{scd}	ρ	$\lvert G_D \rvert$
Country = LT, Join Date $>$ 701, Browser language $\neg$ et-EE	0.9999	0.9782	11
Region = 6	0.9994	-0.1272	10
Region = QUE	0.9992	-0.0788	11
Country = NP	0.9985	0.9630	11
Browser language = es-MX	0.9973	0.1203	7

correlation as follows:

$$\begin{aligned} \widehat{r} &= \frac{\sum (q^i - \bar{q})(g^i - \bar{g})}{\sqrt{\sum (q^i - \bar{q})^2 \sum (g^i - \bar{g})^2}} \\ z' &= \frac{1}{2} \ln \left(\frac{1 + \widehat{r}}{1 - \widehat{r}} \right) \\ z^* &= \frac{z' - z^C}{\sqrt{\frac{1}{|G_D| - 3} + \frac{1}{|G_D^C| - 3}}} \end{aligned} \tag{3.1}$$

Here, $\widehat{r}$ represents the sample correlation, q^i, g^i represent the activity length and course grade of each student, and $\bar{q}, \bar{g}$ represent their average values over the dataset. Equation (3.1) is the Fisher z transformation, z' in the lower equation represents the z' computation on the subgroup and z^C on its complement, and $|G_D|$ represents the number of records covered by subgroup with description D. Under the null hypothesis that the correlation between q and g is the same inside and outside of the subgroup, z^* follows a standard normal distribution. Hence, the value for z^* implies a p-value under this null hypothesis. Leman et al. (Leman et al., 2008) propose to use one minus this p-value as quality measure φ_{scd}: the higher this value is, the more certain we are that the null hypothesis is false and hence exceptional correlations are observed.

Using this quality measure, we conduct the experiment with beam width 20 and search depth 3. In Table 3.6 and Figure 3.11, we list the top-5 subgroups with exceptional quality score, coefficients, and coverage. We can see that some students gain extremely high grades with longer behavior sequence (cf. Figure 3.11b, 3.11e); some students have longer behavior sequence length but lower grades (cf. Figure 3.11c, 3.11d); and for some subgroups, the length of behavior sequences has no obvious correlation with the grades (cf. Figure

3.11f). We can deduce that the efforts that some students spend in the study are not directly correlated with their learning results.

Exceptional Classifier Behavior Analysis Students' behavioral records in MOOCs are sparse, inconsistent and heterogeneous. Learning behavior could be very different in some students comparing with the others. This imbalance increases the difficulty of training a classifier that can perform well on each part of the dataset. This makes it difficult to train a model that is qualified for tasks like dropout prediction and course passing state prediction.

In this section, we investigate whether learning behavior can predict whether or not a student can pass the course. At the same time, we investigate in which parts of the dataset the classifier does not work well. In previous parts, we have presented that EMM can effectively discover exceptional learning behavioral patterns in MOOCs. We will continue using the EMM framework to find where our predictive model does not work well in the dataset. Considering the activities of students in assessments, forum discussions and peer assignments, we formulate the passing state prediction problem as follows:

$$f : \mathcal{X}^i \rightarrow Y^i$$

Our aim is to train a classifier f that can automatically map $\mathcal{X}^i$ to Y^i, where $\mathcal{X}^i$ is a 8-tuple $(s^i, m^i, o^i, c^i, b^i, e^i, h^i, p^i)$ feature vector representing the length of assessment and question sequence (s^i), number of assessment types (m^i), number of question types (o^i), number of correctly answered questions (c), number of asked, answered and liked questions in the forum (b^i, e^i, h^i), and peer review score (p^i), and where Y is the label of passing state: $\{0, 1\}$. We normalize the features into 0 to 1 as the input values.

At first, the classifier is trained on the whole dataset. This model will classify some students correctly and some students wrongly; in any case we find a value of predicted labels $\hat{Y}$. These two binary values Y and $\hat{Y}$ will agree and disagree on some students, and that interaction can be used to capture the quality of the classifier predictions in a single number. We use the f1 score to capture the quality of classifiers:

$$\varphi_{\text{f1}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.2}$$

However, we can perform the exact same computation for a subset of the vectors Y and $\hat{Y}$, for instance the subset induced by a subgroup. Thus, we employ φ_{f1} as a quality measure for EMM.

Table 3.7: Course statistics.

course_name	course_level	complete_number	avg_grades	course_enroll_num	max_grades	min_grades	pass_number
Marketing	I	1141	0.105	4609	1	0.006	52
Design Thinking	I	369	0.167	3483	0.972	0.01	22
IoT	A	8	0.098	241	0.1	0.087	0
System Validation (2)	I	63	0.412	1010	1	0.05	12
Smart IoT	B	905	0.216	6035	1	0.004	100
Computer Architecture	I	913	0.510	7652	1	0.025	299
System Validation (4)	A	17	0.597	985	1	0.071	9
Quantitative Model (1)	I	429	0.395	1807	1	0.007	49
System Validation (3)	A	45	0.418	764	1	0.057	11
Quantitative Model (2)	A	979	0.339	4975	1	0.016	52
System Validation	I	601	0.376	2605	1	0.04	124
Technology	I	258	0.232	3930	1	0.002	34
Embedded Systems	I	549	0.291	3737	1	0.02	67
Software Architecture	A	2710	0.299	10487	1	0.012	331
Real-Time Systems	I	3615	0.203	15123	1	0.006	389
IoT Devices	I	430	0.318	6609	1	0.008	85
Embedded Hardware	I	3943	0.160	19592	1	0.02	128
Open Innovation	I	480	0.137	3150	0.969	0.008	24

Table 3.8: Exceptional classifier behavior for course passing state prediction. Results indicate that the classifier cannot work well on these exceptional subgroups.

D	φ_{f1}	$\lvert G_D \rvert$
Country = OM, Profile language = en-US, Browser language ¬ en-US, Educational status ¬ BACHELOR DEGREE	0.5051	32
Country = OM, Profile language ¬ en-US	0.4058	22
Region = MA, Gender = female, Educational status=COLLEGE NO DEGREE	0.3489	24
Country = OM, Met Payment Condition ¬ True	0.3464	31
Join Date $\leq$ 390, Region ¬ MA	0.3193	28

We conduct the experiment by setting the search depth to 4 and beam width to 10. We engage an (Support Vector Machine) SVM classifier as the predictive model[2], which has 0.85 as f1 score on the whole dataset. In Table 3.8 we list the top-5 subgroups with exceptional behavior. We can see that even though the classifier performs well on the whole dataset, in some subgroups it does not. Particularly for the students described by descriptions like "D: Region = MA, Gender = female, Educational status=COLLEGE NO DEGREE", the classifier performs poorly on the prediction task at hand: the support vector machine has trouble predicting the study success of Massachusets women without a college degree. Hence, this group requires a more sophisticated classifier.

3.4 Conclusion

In this chapter, we study the uncertainty in dependency modeling considering the heterogeneous and high-dimensional interactions between target variables. We develop new quality measures for Exceptional Model Mining with two practical applications: exceptional study behavior analysis and fairness in network representation.

In exceptional learning behavior analysis for MOOC, rather than predicting the success of individual students, which is difficult due to the inherent sparsity, inconsistency, and heterogeneity of the data, EMM specializes in identifying coherent groups that behave differently from the norm. Since the

[2] one may plug in one's preferred classifier; SVM selection is merely meant as an illustration, not an endorsement.

subgroups resulting from EMM come with an easily interpretable definition, *exceptional model mining* allows educators to more effectively channel their remedial activities.

We employ three EMM model classes for different tasks of learning behavior analysis. Experimental results on a real Coursera dataset show that for some students, the dropout rate is very different from the whole dataset, the learning efforts are not always correlated with course grades, and a classifier that performs very well on the whole dataset has trouble on some subpopulations of the data.

For fairness in network representation, we argue that the structural heterogeneity in networks can bias the network representation models across subgroups, which will prevent us from building fair decision making models for downstream tasks like node classification or link prediction. However, the unknown distribution of structural heterogeneity raises new challenges for fairness measurement. Pre-defined groups with sensitive variables are not proper for overcoming the new challenges, and statistical parity with regard to decision variable cannot be helpful for comparing the multi-degree interactions between node representations. We analyze the connections between the structural properties and the node representations in networks. Then we design a framework to compare the node representations learned from subgroups with the node representations learned from the whole data. The differences between them indicate that the structural properties in subgroups are ignored by the network representation model. The higher the difference, the more unfair the model is on those subgroups. The discovery process is automatically guided by a search algorithm defined over the description space, with a quality measure over the learned node representations, called Mean Latent Similarity Discrepancy (MLSD). We evaluate the statistical significance of the discovered subgroups by applying a kernel two-sample test. To validate the effectiveness of our method, we use randomization techniques to generate synthetic datasets with ground truth. This allows us to evaluate the performance of our method quantitatively and qualitatively.

4

Uncertainty in Causal Dependency

"Human reason has this peculiar fate that in one species of its knowledge it is burdened by questions which, as prescribed by the very nature of reason itself, it is not able to ignore, but which, as transcending all its powers, it is also not able to answer."

Critique of Pure Reason ,
Immanuel Kant, 1781.

4.1 Introduction

In this chapter, we introduce a kind of directional dependency between variables, causal dependency. Causal dependency reflects causal relation that determines the generating process between variables. Instead of association dependency, causal dependency is asymmetric in temporal and functional directions.

Definition 4.1.1 (Causal Dependency) Assume we have random variables $X, Y \in \mathbb{R}$. Causal dependency implies a stochastic process that determines the distribution of $Y := f(X) + \epsilon$, where ϵ is the randomness term that adds uncertainty to the value of Y, which is independent to X. Function $f(X)$ represents the deterministic process between X and Y.

However, due to the confounding bias and selection bias in historical data, estimating causal dependency from datasets is challenging. This would bring extra uncertainty to the *exceptional model mining* with causal models as target of interest. Conditional probability may give us an illusion with spurious association between variables. Holland (1986) pointed out that if two variables are correlated with each other, then either there is causal dependency between

them, or there is (are) third part of variable(s) that confounded both of them. Properly handling causal dependency could give us more confidence about the exceptionalities of discovered subgroups.

Learning causal dependency from observational data greatly benefits a variety of domains such as health care, education and sociology. For instance, one could estimate the impact of a new drug to improve the survival rate. In this chapter, we conduct causal inference with observational studies based on the Potential Outcome framework (PO) (Rubin, 2005). The central problem for causal effect inference in PO is dealing with the unobserved counterfactuals and treatment selection bias. The state-of-the-art approaches focus on solving these problems by balancing the treatment and control groups (Sun and Nikolaev, 2016). However, during the learning and balancing process, highly predictive information from the original covariate space might be lost. In order to build more robust estimators, we tackle this information loss problem by presenting a method called **A**dversarial **B**alancing-based representation learning for **C**ausal **E**ffect **I**nference (**ABCEI**), based on the recent advances in representation learning. ABCEI uses adversarial learning to balance the distributions of treatment and control group in the latent representation space, without any assumption on the form of the treatment selection/assignment function. ABCEI preserves useful information for predicting causal effects under the regularization of a mutual information estimator. The experimental results show that ABCEI is robust against treatment selection bias, and matches/outperforms the state-of-the-art approaches. Our experiments show promising results on several datasets, representing different health care domains among others.

4.2 Motivation

Many domains of science require inference of causal effects, including healthcare (Casucci et al., 2017), economics and marketing (LaLonde, 1986, Smith and Todd, 2005), sociology (Morgan and Harding, 2006) and education (Zhao and Heffernan, 2017). For instance, medical scientists must know whether a new medicine is more beneficial for patients; teachers want to know if the teaching plan can be beneficial for students; economists need to evaluate how a policy affects the unemployment rates. Properly estimating causal effects is an important task for machine learning research.

Conducting Randomized Controlled Trials (RCT) can be time-consuming, expensive, or unethical (e.g. for studying the effect of smoking). Hence, ap-

proaches for causal inference from observational data are needed. The core issue of causal effect inference from observational data is confounding: variables might affect both intervention and treatment outcomes. For example, patients with more personal wealth are in a better position to get new medicines, increasing the likelihood that they survive. Inferring causal effect without controlling for confounders will lead to errors. Throughout this chapter, we assume that all the variables in the causal system can be observed and measured, so that the causal effects we are interested can be identifiable from the observational data (Pearl, 2009).

Under the Potential Outcome framework, people usually focus on matching / balancing covariates according to confounders, e.g. based on mutual information (Sun and Nikolaev, 2016) or propensity scores (Dehejia and Wahba, 2002). Average Treatment Effect (ATE) or Average Treatment effect on the Treated (ATT) can be properly estimated after those steps. To account for heterogeneity in subpopulations (Pearl, 2017, Bertsimas et al., 2018), articles about Conditional Average Treatment Effects (CATE) have come out recently (Shalit et al., 2017, Lu et al., 2018). CATE can be estimated by regressing the difference of Individual Treatment Effects (ITEs), which cannot be directly observed from the data, because of the unobservable counterfactuals (Künzel et al., 2019). The main challenges for CATE estimation are two-fold: on the one hand, in observational data, we only know the factual outcome of each unit (treated or untreated), but we will never know the counterfactual outcome; on the other hand, usually the distributions of covariates in treatment and control group are unbalanced (*treatment selection bias*). If we directly employ the standard supervised learning framework to learn the treatment outcome, we will get a biased model suffering from generalization error (Swaminathan and Joachims, 2015b).

4.3 Contributions

To overcome these challenges, we propose a unified framework to encode the input covariates into a latent representation space, and estimate the treatment outcomes with those representations. There are three components on top of the encoder in our model: (1) **mutual information estimation**: an estimator is specified to estimate and maximize the mutual information between representations and covariates; (2) **adversarial balancing**: the encoder plays an adversarial game with a discriminator, trying to fool the discriminator by minimizing the discrepancies between distributions of representations from the

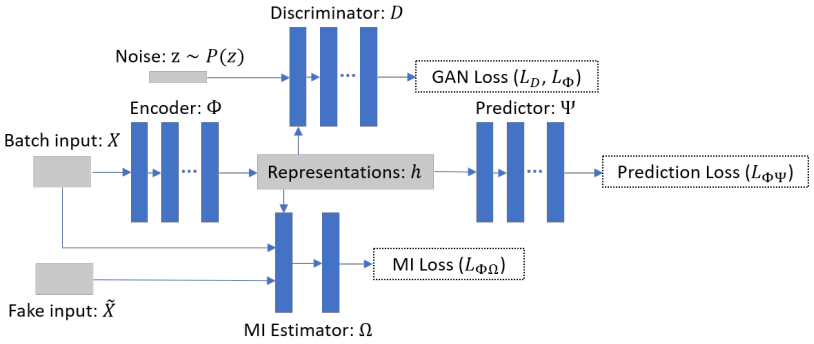


Figure 4.1: Deep neural network architecture of ABCEI for causal effect inference.

treatment and control group; (3) **treatment outcome prediction**: a predictor over latent space is employed to estimate the treatment outcomes. By jointly optimizing the three components via back propagation, we can get a robust estimator on causal effects. The overarching architecture of our framework is shown in Figure 4.1. As a summary, our main contributions are:

1. We propose a novel model: **A**dversarial **B**alancing-based representation learning for **C**ausal **E**ffect **I**nference (**ABCEI**) with observational data. ABCEI addresses information loss and selection bias by learning highly informative and balanced representations in latent space.
2. A neural network encoder is constrained by a mutual information estimator to minimize the information loss between representations and the input covariates, which preserves highly predictive information for causal effect inference.
3. We employ an adversarial learning method to balance representations between treatment and control groups, which deals with the selection bias problem without any assumption on the form of the treatment selection function, unlike, e.g., the propensity score method.
4. We conduct various experiments on synthetic and real-world datasets. ABCEI outperforms most of the state-of-the-art methods on benchmark datasets. We show that ABCEI is robust against different experimental settings. By supporting mini-batch, ABCEI can be applied on large-scale datasets.

4.4 Related Work

Work on causality learning falls into two categories: causal inference and causal discovery (Mooij et al., 2016). In the branch of causal inference, three kinds of data are used: data from Randomized Controlled Trials (RCT), observational data for which all the (potential) confounders can be observed, and observational data with unobserved confounders. A branch of research with RCT datasets focuses on identification of heterogeneous treatment effects. Both machine learning (Lamont et al., 2018, Taddy et al., 2016) and optimization (Bertsimas et al., 2018) approaches are applied. Due to the difficulties of obtaining RCT datasets, observational studies become an alternative. Removing confounding is a core issue in causal inference with observational data. Confounding bias, selection bias and missing data are three main problems for causal inference with observational data. Some research estimates population causal effects with an instrumental variable (Bareinboim and Pearl, 2012); some research uses latent variable models to simultaneously discover hidden confounders and estimate causal effects (Louizos et al., 2017), which is robust against hidden confounding; some research focus on the recoverability in the presence of selection bias (Correa et al., 2019). In this chapter, we assume that all the studied variables can be measured, which satisfies the strong ignorability assumption (Rosenbaum and Rubin, 1983).

In this branch, one of the core issues to deal with is the mismatch between treatment and control groups. From the view of balancing, there are three ways. The first and classical way of balancing is referred to as matching (Ho et al., 2011): a control group is selected in order to maximize the similarity between the empirical covariate distributions in the treatment and control group. Mahalanobis distance and propensity score matching methods are proposed for population causal effect inference (Rubin, 2001, Diamond and Sekhon, 2013). An information theory-driven approach is proposed by using mutual information as the similarity measure (Sun and Nikolaev, 2016). In the second way of balancing, the Inverse Propensity Score (IPS) method is proposed based on the variants of importance sampling (Sugiyama and Krauledat, 2007, Jiang and Li, 2016). The IPS is used to reweigh each unit sample to learn the counterfactuals, which is akin to counterfactual learning from logged bandit feedback (Swaminathan and Joachims, 2015b,a). In the third way, methods from representation learning are used to transform covariates from the original space into a latent representation space (Li and Fu, 2017). The representations are used as the input of predictors for individual and population causal effect inference. One study reported on use of a single neural network with the con-

catenation of representations and treatment variable as the input (Johansson et al., 2016). Separate models were trained for different treatments associated with a probabilistic integral metric to bound the generalization errors in (Shalit et al., 2017). Hard samples to preserve local similarity during balancing process were used in (Yao et al., 2018). Our methods are most similar to these third-way methods. The main difference between ABCEI and the existing approaches is that except balancing, we address the information loss problem by simultaneously estimating and maximizing the mutual information between latent representations and the input covariates.

From the technical viewpoint, our method lies into the field of representation learning. The main aim of learning representations is to obtain useful information from original data for downstream tasks like building predictors or classifiers. From Principal Components Analysis (PCA) (Smith, 2002) to autoencoders (Vincent et al., 2008), many approaches account for learning representations. A proper way to evaluate the quality of learned representations is to measure the reconstruction error (Kingma and Welling, 2013). Specifically, reconstruction error is shown to be minimized by maximizing mutual information between input and the learned representations when their joint distributions for the encoder and decoder are matched (Belghazi et al., 2018). As a consequence, maximizing mutual information minimizes the information loss and the expected reconstruction error. We adopt this approach to regularize the encoder to preserve useful information for prediction tasks. However, in continuous and high-dimensional spaces, accurately computing MI is quite difficult. KL-divergence (Donsker and Varadhan, 1983) and Jensen-Shannon-divergence (JSD) (Nowozin et al., 2016) based methods are introduced for approximating mutual information with neural networks. We follow this way to build the neural network estimator for MI estimation.

More and more machine learning methods are employed for causal inference. For instance, Bayesian additive regression trees and Random forests were employed to estimate causal effects in (Sparapani et al., 2016) and (Wager and Athey, 2017) respectively. Some research discusses how domain adaptation (Daume III and Marcu, 2006) and Generative Adversarial Networks (GAN) (Goodfellow, 2016) can be used for causal inference by generating balanced weights for unit samples (Ozery-Flato et al., 2018, Kuang et al., 2018). Fitting a model only with observed factual data by using the GAN framework, which is suitable for any number of treatments was proposed in (Yoon et al., 2018). The main difference between ABCEI and those methods is that we use adversarial learning to balance distributions of treatment group and control group in the latent representation space.

ABCEI does not need prior knowledge about treatment assignment. By following the design of Wasserstein GAN (Gulrajani et al., 2017), our adversarial balancing can make the encoder generate more similar distributions for treatment and control group. Another advantage of our method is that we account for the information loss problem by using a mutual information estimator to regularize the encoder. The mutual information estimator uses a neural network to simultaneously approximate and minimize the information loss, which persuades the encoder to learn representations preserving highly predictive information. Based on those advantages, the two components – mutual information estimator and adversarial balancing – combined together allow us to find the proper predictor for causal effect inference.

4.5 Methodology

4.5.1 Preliminaries

In order to properly handle treatment selection bias and counterfactuals, causal effect estimation must solve two central problems: balancing covariates and specifying the outcome model. Recent methods in causal inference tackle one or both of these problems. (Yao et al., 2018) propose to use hard samples, to preserve local similarity information from covariate space to latent representation space. The hard sample mining process is highly dependent on the propensity score model, which is not robust when the propensity score model is misspecified. (Imai and Ratkovic, 2014, Ning et al., 2018) propose estimators which are robust even when the propensity score model is not correctly specified. (Kallus, 2018a,b, Ozery-Flato et al., 2018) propose to generate balanced weights for data samples to minimize a selected imbalance measure in covariate space. (Shalit et al., 2017) propose to derive upper bounds on the estimation error by considering both covariate balancing and potential outcomes. Highly predictive information might be lost in the reweighing or balancing processes of these methods.

To address these problems, we propose a framework (cf. Figure 4.1), which generates balanced representations preserving highly predictive information in latent space without considering propensity scores. We design a two-player adversarial game, between an encoder that transforms covariates to latent representations and a discriminator which distinguishes representations from control and treatment group. Unlike in the classical GAN framework, here, the

'true distribution' (latent representations of the control group[1]) in this game also must be generated by the encoder. On the other hand, to prevent losing useful information during the balancing process, we use a mutual information estimator to constrain the encoder to preserve highly predictive information (Hjelm et al., 2018). The outcome data are also considered in this unified framework to specify the causal effect predictor.

Problem Setup Assume an observational dataset $\{X, T, Y\}$, with covariate matrix $X \in \mathbb{R}^{n \times k}$, binary treatment vector $T \in \{0, 1\}^n$, and treatment outcome vector $Y \in \mathbb{R}^n$. Here, n denotes the number of observed units, and k denotes the number of covariates in the dataset. For each unit u, we have k covariates $x_1, \ldots, x_k$, associated with one treatment variable $t \in \{0, 1\}$ and one treatment outcome y. According to the Rubin-Neyman causal model (Rubin, 2005), two potential outcomes y_0, y_1 exist for treatments $\{0, 1\}$, respectively. We call y_t the *factual outcome*, denoted by y_f, and y_{1-t} the *counterfactual outcome*, denoted by y_{cf}. Assuming there is a joint distribution $P(x, t, y_0, y_1)$, we make the following assumptions:

Assumption 4.5.1 Conditioned on x, the potential outcomes y_0, y_1 are independent of t, which can be stated as: $(y_0, y_1) \perp\!\!\!\perp t|x$.

Assumption 4.5.2 For all sets of covariates and for all treatments, the probability of treatment assignment will always be strictly larger than 0 and strictly smaller than 1, which can be expressed as: $0 < P(t|x) < 1, \forall t$ and $\forall x$.

Assumption 4.5.1 indicates that all the confounders are observed, i.e., *no unmeasured confounder is present*. Assumption 4.5.2 allows us to estimate the CATE for any x in the covariate space. Under these assumptions, we can formalize the definition of CATE (Shalit et al., 2017) for our task:

Definition 4.5.1 The Conditional Average Treatment Effect (CATE), for unit u is: $CATE(u) := \mathbb{E}[\, y_1 \mid x^u \,] - \mathbb{E}[\, y_0 \mid x^u \,]$.

We can now define the Average Treatment Effect (ATE) and the Average Treatment effect on the Treated (ATT) as:

$$ATE := \mathbb{E}[\, CATE(u) \,] \quad ATT := \mathbb{E}[\, CATE(u) \mid t = 1 \,].$$

[1] our method supports representations of either treatment/control group or both as 'true distribution'.

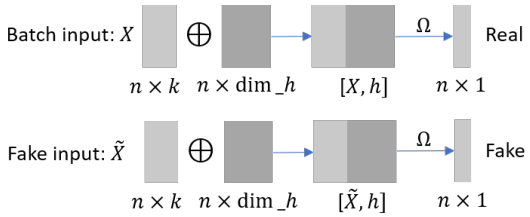


Figure 4.2: MI estimator between covariates and latent representations.

Because the joint distribution $P(x, t, y_0, y_1)$ is unknown, we can only try to estimate $CATE(u)$ from observational data. A function over the covariate space $\mathcal{X}$ can be defined as $f : \mathcal{X} \times \{0, 1\} \rightarrow \mathcal{Y}$. The estimate of $CATE(u)$ can now be defined:

Definition 4.5.2 Given an observational dataset $\{X, T, Y\}$ and a function f, for unit u, the estimate of $CATE(u)$ is:

$$\widehat{CATE}(u) = f(x^u, 1) - f(x^u, 0).$$

In order to properly accomplish the task of CATE estimation, we need to find an optimal function over the covariate space for both systems ($t = 1$ and $t = 0$).

4.5.2 Neural Network Framework for Counterfactual Prediction

In order to overcome the challenges in CATE estimation, we build our model on recent advances in representation learning. We propose to define a function $\Phi : \mathcal{X} \rightarrow \mathcal{H}$, and a function $\Psi : \mathcal{H} \rightarrow \mathcal{Y}$. Then we have $\widehat{Y_T} = f(X, T) = \Psi(\Phi(X), T) = \Psi(h, T)$. Instead of directly estimating the treatment outcome conditioned on covariates, we firstly use an encoder to learn latent representations of covariates. We simultaneously learn latent representations and estimate the treatment outcome. However, the function f would still suffer from information loss and treatment selection bias, unless we constrain the encoder Φ to learn balanced representations while preserving useful information.

Mutual Information Estimation Consider the information loss when transforming covariates into latent space. The non-linear statistical dependencies between variables can be acquired by mutual information (MI) (Kinney and

Atwal, 2014). Thus we use MI between latent representations and original covariates as a measure to account for information loss:

$$I(X;h) = \int_{\mathcal{X}} \int_{\mathcal{H}} P(x,h) \log \left(\frac{P(x,h)}{P(x)P(h)} \right) \mathrm{d}h \, \mathrm{d}x.$$

We denote the joint distribution between covariates and representations by $\mathbb{P}_{Xh}$ and the product of marginals by $\mathbb{P}_X \otimes \mathbb{P}_h$. From the viewpoint of Shannon information theory, mutual information can be represented as Kullback-Leibler (KL) divergence:

$$I(X;h) := H(X) - H(X|h) := D_{KL}(\mathbb{P}_{Xh}||\mathbb{P}_X \otimes \mathbb{P}_h),$$

It is hard to compute MI in continuous and high-dimensional spaces, but one can capture a lower bound of MI with the Donsker-Varadhan representation of KL-divergence (Donsker and Varadhan, 1983):

Theorem 4.5.1 (Donsker-Varadhan)

$$D_{KL}(\mathbb{P}_{Xh}||\mathbb{P}_X \otimes \mathbb{P}_h) = \sup_{\Omega \in \mathcal{C}} \mathbb{E}_{\mathbb{P}_{Xh}}[\Omega(x,h)] - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_h} \left[e^{\Omega(x,h)} \right].$$

Here, $\mathcal{C}$ denotes the set of unconstrained functions Ω.

Proof. Given a fixed function Ω, we can define distribution G by:

$$\mathrm{d}G = \frac{e^{\Omega(Z)} \mathrm{d}Q}{\int_{\mathcal{Z}} e^{\Omega(Z)} \mathrm{d}Q}$$

Equivalently, we have:

$$\mathrm{d}G = e^{(\Omega(Z) - S)} \mathrm{d}Q \ , \qquad S = \log \mathbb{E}_Q \left[e^{\Omega(Z)} \right]$$

Then by construction, we have:

$$\begin{aligned}
&\mathbb{E}_P[\Omega(Z)] - \log \mathbb{E}_Q \left[e^{\Omega(Z)} \right] \\
&= \mathbb{E}_P[\Omega(Z)] - S \\
&= \mathbb{E}_P \left[\log \frac{\mathrm{d}G}{\mathrm{d}Q} \right] \\
&= \mathbb{E}_P \left[\log \frac{\mathrm{d}P \mathrm{d}G}{\mathrm{d}Q \mathrm{d}P} \right] \\
&= \mathbb{E}_P \left[\log \frac{\mathrm{d}P}{\mathrm{d}Q} - \log \frac{\mathrm{d}P}{\mathrm{d}G} \right] \\
&= D_{KL}(P||Q) - D_{KL}(P||G) \\
&\leq D_{KL}(P||Q)
\end{aligned}$$

When distribution G is equal to P, this bound is tight. ■

Inspired by Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018), we propose to establish a neural network estimator for MI. Specifically, let Ω be a function: $\mathcal{X} \times \mathcal{H} \to \mathbb{R}$ parametrized by a deep neural network, we have:

$$\begin{aligned} I(X;h) &:= D_{KL}\left(\mathbb{P}_{Xh} || \mathbb{P}_X \otimes \mathbb{P}_h\right) \geq \hat{I}_\Omega(X;h) \\ &:= \mathbb{E}_{\mathbb{P}_{Xh}}[\Omega(x,h)] - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_h}\left[e^{\Omega(x,h)}\right]. \end{aligned} \tag{4.1}$$

By distinguishing the joint distribution and the product of marginals, the estimator Ω approximates the MI with arbitrary precision. In practice, as shown in Figure 4.2, we concatenate the input covariates X with representations h one by one to create positive samples (as samples from the true joint distribution). Then, we randomly shuffle X on the batch axis to create fake input covariates $\tilde{X}$. Representations h are concatenated with fake input $\tilde{X}$ to create negative samples (as samples from the product of marginals). From Equation (4.1) we can derive the loss function for the MI estimator:

$$L_{\Phi\Omega} = -\mathbb{E}_{x \sim X}\left[\Omega\left(x,h\right)\right] + \log \mathbb{E}_{x \sim \tilde{X}}\left[e^{\Omega(x,h)}\right].$$

Information loss can be diminished by simultaneously optimizing the encoder Φ and the MI estimator Ω to minimize $L_{\Phi\Omega}$ iteratively via gradient descent.

Adversarial Balancing The representations of treatment and control groups are denoted by $h(t=1)$ and $h(t=0)$, corresponding to the input covariate groups $X(t=1)$ and $X(t=0)$. The discrepancy between distributions of the treatment and control groups is an urgent problem in need of a solution. To decrease this discrepancy, we propose an adversarial learning method to constrain the encoder to learn treatment and control representations that are balanced distributions. We build an adversarial game between a discriminator D and the encoder Φ, inspired by the framework of GAN (Goodfellow et al., 2014). In the classical GAN framework, a source of noise is mapped to a generated image by a generator. A discriminator is trained to distinguish whether an input sample is from true or synthetic image distribution generated by the generator. The aim of classical GAN is training a reliable discriminator to distinguish fake and real images, and using the discriminator to train a generator to generate images by fooling the discriminator.

In our adversarial game: (1) we draw a noise vector $z \sim P(z)$ which has the same length as the latent representations, where $P(z)$ can be a spherical

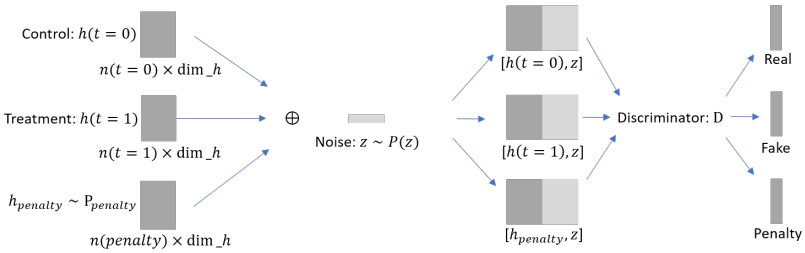


Figure 4.3: Adversarial learning structure for representation balancing.

Gaussian distribution or a Uniform distribution; (2) we separate representation by treatment assignment, and form two distributions: $P_{h(t=1)}$ and $P_{h(t=0)}$; (3) we train a discriminator D to distinguish concatenated vectors from treatment and control group ($[z, h(t=1)]$ and $[z, h(t=0)]$); (4) we optimize the encoder Φ to generate balanced representations to fool the discriminator.

According to the architecture of ABCEI, the encoder is associated with the MI estimator Ω, treatment outcome predictor Ψ and adversarial discriminator D. This means that the training process is iteratively adjusting each of the components. The instability of GAN training will become serious in this context. To stabilize the training of GAN, we propose to use the framework of Wasserstein GAN with gradient penalty (Gulrajani et al., 2017). By removing the sigmoid layer and applying the gradient penalty to the data between the distributions of treatment and control groups, we can find a function D which satisfies the 1-Lipschitz inequality:

$$\left|\left|D\left(x^1\right) - D\left(x^2\right)\right|\right| \leq \left|\left|x^1 - x^2\right|\right|.$$

We can write down the form of our adversarial game:

$$\min_{\Phi} \max_{D} \mathbb{E}_{h \sim P_{h(t=0)}}[D([z,h])] - \mathbb{E}_{h \sim P_{h(t=1)}}[D([z,h])] - \beta\, \mathbb{E}_{h \sim P_{\text{penalty}}}\left[\left(||\nabla_{[z,h]} D([z,h])||_2 - 1\right)^2\right],$$

where P_{penalty} is the distribution acquired by uniformly sampling along the straight lines between pairs of samples from $P_{h(t=0)}$ and $P_{h(t=1)}$. The adversarial learning process is in Figure 4.3.

This ensures the encoder Φ to be smoothly trained to generate balanced representations. We can write down the training objective for discriminator

and encoder, respectively:

$$\begin{aligned} L_D =& -\mathbb{E}_{h\sim P_{h(t=0)}}[D([z,h])] + \mathbb{E}_{h\sim P_{h(t=1)}}[D([z,h])] \\ &+ \beta\,\mathbb{E}_{h\sim P_{\text{penalty}}}\left[(||\nabla_{[z,h]}D([z,h])||_2 - 1)^2\right], \\ L_\Phi =& \mathbb{E}_{h\sim P_{h(t=0)}}[D([z,h])] - \mathbb{E}_{h\sim P_{h(t=1)}}[D([z,h])]. \end{aligned}$$

Treatment Outcome Prediction The final step for CATE estimation is to predict the treatment outcomes with learned representations. We establish a neural network predictor, which takes latent representations and treatment assignments of units as the input, to conduct outcome prediction: $\widehat{y_t} = \Psi(h,t)$. We can write down the loss function of the training objective as:

$$L_{\Phi\Psi} = \mathbb{E}_{(h,t,y_t)\sim\{h,T,Y_T\}}\left[\left(\Psi\left(h,t\right) - y_t\right)^2\right] + \lambda\, R(\Psi).$$

Here, R is a regularization on Ψ for the model complexity.

4.5.3 Learning Optimization

With respect to the architecture in Figure 4.1, we minimize $L_{\Phi\Omega}$, L_Φ, and $L_{\Phi\Psi}$, respectively, to iteratively optimize parameters in the global model. The optimization steps are handled with the stochastic method Adam (Kingma and Ba, 2014), training the model within Algorithm 2. Optimization details and computational complexity analysis are given in the supplementary material.

4.5.4 Training Details

The implementation of our method is based on Python and Tensorflow (Abadi et al., 2016). All the experiments in this chapter are conducted on a cluster with 1x Intel Xeon E5 2.2GHz CPU, 4x Nvidia Tesla V100 GPU and 256GB RAM.

We adopt Exponential Linear Unit (ELU) (Clevert et al., 2015) as the non-linear activation function if there is no specification. We employ various numbers of fully-connected hidden layers with various sizes across networks: four layers with size 200 for the encoder network; two layers with size 200 for the mutual information estimator network; three layers with size 200 for the discriminator network; and finally, three layers with size 100 for the predictor network, following the structure of TARnet (Shalit et al., 2017). The gradient penalty weight β is set to 10.0, and the regularization weight is set to 0.0001.

Algorithm 2 ABCEI

Input: Observational dataset $\{X, T, Y\}$; loss function $L_{\Phi\Omega}$, L_{Φ} and $L_{\Phi\Psi}$, L_D; Neural Networks Φ, Ω, D, Ψ; parameters Θ_Φ, Θ_Ω, Θ_D, Θ_Ψ
repeat
  Draw mini-batch $\{X_b, T_b, Y_b\} \subset \{X, T, Y\}$
  Compute representations $h = \Phi(X_b)$
  Draw fake input $\tilde{X}_b \sim \tilde{\mathbb{P}}$
  Draw noise $z \sim \mathcal{N}(0, I)$
  Set Θ_Φ, $\Theta_\Omega \leftarrow \text{Adam}(L_{\Phi\Omega}(X_b, \tilde{X}_b, h), \Theta_\Phi, \Theta_\Omega)$
  for $i = 1$ to 3 **do**
    Set $\Theta_D \leftarrow \text{Adam}(L_D(h, z, T_b), \Theta_D)$
  end for
  Set $\Theta_\Phi \leftarrow \text{Adam}(L_\Phi(h, z, T_b), \Theta_\Phi)$
  Set Θ_Φ, $\Theta_\Psi \leftarrow \text{Adam}(L_{\Phi\Psi}(h, T_b, Y_b), \Theta_\Phi, \Theta_\Psi)$
until convergence

In the training step, firstly we minimize $L_{\Phi\Omega}$ by simultaneously optimizing Φ and Ω with one-step gradient descent. Then the representations h are passed to the discriminator to minimize L_D by optimizing D with 3-step gradient descent, in order to find a stable discriminator. Next, we use discriminator D to train encoder Φ by minimizing L_Φ with one-step gradient descent. Finally, encoder Φ and predictor Ψ are optimized simultaneously by minimizing $L_{\Phi\Psi}$.

4.5.5 Hyper-parameter Optimization

Due to the reason that we cannot observe counterfactuals in observational datasets, standard cross-validation methods are not feasible. We follow the hyper-parameter optimization criterion in (Shalit et al., 2017), with an early stopping with regard to the lower bound on the validation set. Detail search space of hyper-parameter is demonstrated in Table 4.1. The optimal hyper-parameter settings for each benchmark dataset is demonstrated in Table 4.2.

4.5.6 Computational Complexity

Assuming the size of mini-batch is n, number of epochs is m, the computational complexity of our model is $\mathcal{O}(n * m * ((\Phi_h - 1)\Phi_w^2 + (\Omega_h - 1)\Omega_w^2 + (D_h - 1)D_w^2 + (\Psi_h - 1)\Psi_w^2))$. Here $\Phi_h, \Omega_h, D_h, \Psi_h$ indicates the number of layers and $\Phi_w, \Omega_w, D_w, \Psi_w$ indicates number of neurons in each layer in

Table 4.1: Search space of hyper-parameter

Hyper-parameter	Range
λ	$1e$-3,$1e$-4,$5e$-5
β	1.0,5.0,10.0,15.0
Optimizer	RMSProp, Adam
Depth of encoder layers	$1, 2, 3, 4, 5, 6$
Depth of discriminator layers	$1, 2, 3, 4, 5, 6$
Depth of predictor layers	$1, 2, 3, 4, 5, 6$
Dimension of encoder layers	$50, 100, 200, 300, 500$
Dimension of discriminator layers	$50, 100, 200, 300, 500$
Dimension of MI estimator layers	$50, 100, 200, 300, 500$
Dimension of predictor layers	$50, 100, 200, 300, 500$
Batch size	$65, 80, 100, 200, 300, 500$

Neural Network Φ, Ω, D, Ψ.

4.6 Experiments

There are two ways to validate and test the performance of causal inference methods: the one is to use simulated or semi-simulated treatment outcomes, e.g., dataset IHDP (Hill, 2011); the other is to use RCT datasets and add a non-randomized component to generate imbalanced datasets, e.g., dataset Jobs (LaLonde, 1986, Smith and Todd, 2005). We designed experiments along both paths for evaluating our method. The four benchmark datasets IHDP, Jobs, Twins (Louizos et al., 2017) and ACIC (Dorie et al., 2019) are used. For IHDP, Jobs, Twins and ACIC, the experimental results are averaged over $1000, 100, 100, 7700$ train/validation/test sets respectively with split sizes $60\%/30\%/10\%$.

4.6.1 Details of Datasets

IHDP The *Infant Health and Development Program* (IHDP) studies the impact of specialist home visits on future cognitive test scores. Covariates in the semi-simulated dataset are collected from a real-world randomized experiment. The treatment selection bias is created by removing a subset of the treatment group. We use the setting 'A' in (Dorie, 2016) to simulate treatment outcomes. This dataset includes 747 units (608 control and 139 treated) with

Table 4.2: Optimal hyper-parameter for each benchmark dataset

Hyper-parameters	Datasets			
	IHDP	Jobs	Twins	ACIC
λ	$1e$-4	$1e$-4	$1e$-4	$1e$-4
β	10.0	10.0	10.0	10.0
Optimizer	Adam	Adam	Adam	Adam
Depth of encoder layers	4	5	5	4
Depth of discriminator layers	3	3	3	3
Depth of predictor layers	3	3	3	3
Dimension of encoder layers	200	200	300	200
Dimension of discriminator layers	200	200	200	200
Dimension of MI estimator layers	200	200	200	200
Dimension of predictor layers	100	100	200	100
Batch size	65	100	300	100

25 covariates associated with each unit.

Jobs The *Jobs* dataset (LaLonde, 1986, Smith and Todd, 2005) studies the effect of job training on the employment status. It consists of a non-randomized component from observational studies and a randomized component based on the National Supported Work program. The randomized component includes 722 units (425 control and 297 treated) with seven covariates, and the non-randomized component (PSID comparison group) includes 2490 control units.

Twins The *Twins* dataset is created based on the "Linked Birth / Infant Death Cohort Data" by NBER [2]. Inspired by (Almond et al., 2005), we employ a matching algorithm to select twin births in the USA between 1989-1991. By doing this, we get units associated with 43 covariates including education, age, race of parents, birth place, marital status of mother, the month in which pregnancy prenatal care began, total number of prenatal visits and other variables indicating demographic and health conditions. We only select twins that have the same gender who both weigh less than $2000g$. For the treatment variable, we use $t = 0$ indicating the lighter twin and $t = 1$ indicating the heavier twin. We take the mortality of each twin in their first year of life as the treatment outcome, inspired by (Louizos et al., 2017). Finally, we have a dataset con-

[2] `https://nber.org/data/linked-birth-infant-death-data-vital-statistics-dat`
`html`

sisting of 12,828 pairs of twins whose mortality rate is 19.02% for the lighter twin and 16.54% for the heavier twin. Hence, we have observational treatment outcomes for both treatments. In order to simulate the selection bias, we selectively choose one of the twins to observe with regard to the covariates associated with each unit as follows: $t|x \sim \text{Bernoulli}(\sigma(w^T x + n))$, where $w^T \sim \mathcal{N}(0, 0.1 \cdot I)$ and $n \sim \mathcal{N}(1, 0.1)$.

ACIC The *Atlantic Causal Inference Conference* (ACIC) (Dorie et al., 2019) is derived from real-world data with 4802 observations using 58 covariates. There are 77 datasets which are simulated with different treatment selection and outcome functions. Each dataset is generated with 100 random replications independently. In this benchmark, different settings like degrees of non-linearity, treatment selection bias and magnitude of treatment outcome are considered.

4.6.2 Evaluation Metrics

Since the ground truth CATE for the IHDP dataset is known, we can employ Precision in Estimation of Heterogeneous Effect (PEHE) (Hill, 2011), as the evaluation metric of CATE estimation:

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{u=1}^{n}((\mathbb{E}[y_1|x^u] - \mathbb{E}[y_0|x^u]) - (f(x^u, 1) - f(x^u, 0)))^2.$$

Subsequently, we can evaluate the precision of ATE estimation based on estimated CATE. For the Jobs dataset, because we only know parts of the ground truth (the randomized component), we cannot evaluate the performance of ATE estimation. Following (Shalit et al., 2017), we evaluate the precision of ATT estimation and policy risk estimation, where

$$R_{pol}(\pi) = 1 - [\mathbb{E}(y_1|\pi(x^u) = 1) \cdot P(\pi = 1) + \mathbb{E}(y_0|\pi(x^u) = 0) \cdot P(\pi = 0)].$$

In this chapter, we consider $\pi(x^u) = 1$ when $f(x^u, 1) - f(x^u, 0) > 0$. For the Twins dataset, because we only know the observed treatment outcome for each unit, we follow (Louizos et al., 2017) using Area Under ROC (Receiver Operating Characteristic) Curve (AUC) as the evaluation metric. For ACIC dataset, we follow (Ozery-Flato et al., 2018) to use RMSE ATE as performance metric.

4.6.3 Baseline Methods

We compare with the following baselines: least square regression using treatment as a feature (**OLS/LR$_1$**); separate least square regressions for each treatment (**OLS/LR$_2$**); balancing linear regression (**BLR**) and balancing neural network (**BNN**) (Johansson et al., 2016); k-nearest neighbor (**k-NN**) (Crump et al., 2008); Bayesian additive regression trees (**BART**) (Sparapani et al., 2016); random forests (**RF**) (Breiman, 2001); causal forests (**CF**) (Wager and Athey, 2017); treatment-agnostic representation networks (**TARNet**) and counterfactual regression with Wasserstein distance (**CFR-Wass**) (Shalit et al., 2017); causal effect variational autoencoders (**CEVAE**) (Louizos et al., 2017); local similarity preserved individual treatment effect (**SITE**) (Yao et al., 2018); MMD measure using RBF kernel (**MMD-V1, MMD-V2**) (Kallus, 2018b,a); Adversarial balancing with cross-validation procedure (**ADV-LR/SVM/MLP**) (Ozery-Flato et al., 2018). We show the quantitative comparison between our method and the state-of-the-art baselines. Experimental results (in-sample and out-of-sample) on IHDP, Jobs and Twins datasets are reported. Specifically, we use ABCEI* to represent our model without the mutual information estimation component, and ABCEI** to represent our model without the adversarial learning component.

4.6.4 Results

Experimental results are shown in Tables 4.3 and 4.4. It would be unsound to report statistical test results over the results reported in these tables; due to varying (un-)availability of ground truth, we must resort to reporting varying evaluation measures per dataset, over which it would not be appropriate to aggregate in a single statistical hypothesis test. However, one can see that ABCEI performs best in ten out of twelve cases, not only by the best number in the column, but often also by a non-overlapping empirical confidence interval with that of the best competitor (cf. reported standard deviations). This provides evidence that ABCEI is a substantial improvement over the state of the art.

Due to the existence of treatment selection bias, regression based methods suffer from high generalization error. Nearest neighbor based methods consider unit similarity to overcome selection bias, but cannot achieve balance globally. Recent advances in representation learning bring improvements in causal effect estimation. Unlike CFR-Wass, BNN, and SITE, ABCEI considers information loss and balancing problems. The mutual information es-

Table 4.3: In-sample and out-of-sample results with mean and standard errors on the IHDP and Jobs dataset (lower = better).

Methods	IHDP			
	In-sample		Out-sample	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
OLS/LR_1	$5.8 \pm .3$	$.73 \pm .04$	$5.8 \pm .3$	$.94 \pm .06$
OLS/LR_2	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$5.8 \pm .3$	$.93 \pm .05$
BART	$2.1 \pm .1$	$.23 \pm .01$	$2.3 \pm .1$	$.34 \pm .02$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$4.1 \pm .2$	$.79 \pm .05$
RF	$4.2 \pm .2$	$.73 \pm .05$	$6.6 \pm .3$	$.96 \pm .06$
CF	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$
TARNet	$.88 \pm .0$	$.26 \pm .01$	$.95 \pm .0$	$.28 \pm .01$
CFR-Wass	$.71 \pm .0$	$.25 \pm .01$	$.76 \pm .0$	$.27 \pm .01$
CEVAE	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$
SITE	$\mathbf{.69 \pm .0}$	$.22 \pm .01$	$.75 \pm .0$	$.24 \pm .01$
ABCEI*	$.74 \pm .0$	$.12 \pm .01$	$.78 \pm .0$	$.11 \pm .01$
ABCEI**	$.81 \pm .1$	$.18 \pm .03$	$.89 \pm .1$	$.16 \pm .02$
ABCEI	$.71 \pm .0$	$\mathbf{.09 \pm .01}$	$\mathbf{.73 \pm .0}$	$\mathbf{.09 \pm .01}$
Methods	Jobs			
	In-sample		Out-sample	
	R_{pol}	ϵ_{ATT}	R_{pol}	ϵ_{ATT}
OLS/LR_1	$.22 \pm .0$	$\mathbf{.01 \pm .00}$	$.23 \pm .0$	$.08 \pm .04$
OLS/LR_2	$.21 \pm .0$	$.01 \pm .01$	$.24 \pm .0$	$.08 \pm .03$
BLR	$.22 \pm .0$	$.01 \pm .01$	$.25 \pm .0$	$.08 \pm .03$
BART	$.23 \pm .0$	$.02 \pm .00$	$.25 \pm .0$	$.08 \pm .03$
k-NN	$.23 \pm .0$	$.02 \pm .01$	$.26 \pm .0$	$.13 \pm .05$
RF	$.23 \pm .0$	$.03 \pm .01$	$.28 \pm .0$	$.09 \pm .04$
CF	$.19 \pm .0$	$.03 \pm .01$	$.20 \pm .0$	$.07 \pm .03$
BNN	$.20 \pm .0$	$.04 \pm .01$	$.24 \pm .0$	$.09 \pm .04$
TARNet	$.17 \pm .0$	$.05 \pm .02$	$.21 \pm .0$	$.11 \pm .04$
CFR-Wass	$.17 \pm .0$	$.04 \pm .01$	$.21 \pm .0$	$.08 \pm .03$
CEVAE	$.15 \pm .0$	$.02 \pm .01$	$.26 \pm .1$	$.03 \pm .01$
SITE	$.17 \pm .0$	$.04 \pm .01$	$.21 \pm .0$	$.09 \pm .03$
ABCEI*	$.14 \pm .0$	$.04 \pm .01$	$.18 \pm .0$	$.04 \pm .01$
ABCEI**	$.15 \pm .0$	$.05 \pm .01$	$.19 \pm .0$	$.04 \pm .01$
ABCEI	$\mathbf{.13 \pm .0}$	$.02 \pm .01$	$\mathbf{.17 \pm .0}$	$\mathbf{.03 \pm .01}$

Table 4.4: In-sample and out-of-sample results with mean and standard errors on the Twins dataset (AUC: higher = better, ϵ_{ATE}: lower = better).

Methods	In-sample		Out-sample	
	AUC	ϵ_{ATE}	AUC	ϵ_{ATE}
OLS/LR_1	$.660 \pm .005$	$.004 \pm .003$	$.500 \pm .028$	$.007 \pm .006$
OLS/LR_2	$.660 \pm .004$	$.004 \pm .003$	$.500 \pm .016$	$.007 \pm .006$
BLR	$.611 \pm .009$	$.006 \pm .004$	$.510 \pm .018$	$.033 \pm .009$
BART	$.506 \pm .014$	$.121 \pm .024$	$.500 \pm .011$	$.127 \pm .024$
k-NN	$.609 \pm .010$	$.003 \pm .002$	$.492 \pm .012$	$.005 \pm .004$
BNN	$.690 \pm .008$	$.006 \pm .003$	$.676 \pm .008$	$.020 \pm .007$
TARNet	$.849 \pm .002$	$.011 \pm .002$	$.840 \pm .006$	$.015 \pm .002$
CFR-Wass	$.850 \pm .002$	$.011 \pm .002$	$.842 \pm .005$	$.028 \pm .003$
CEVAE	$.845 \pm .003$	$.022 \pm .002$	$.841 \pm .004$	$.032 \pm .003$
SITE	$.862 \pm .002$	$.016 \pm .001$	$.853 \pm .006$	$.020 \pm .002$
ABCEI*	$.861 \pm .001$	$.005 \pm .001$	$.851 \pm .001$	$.006 \pm .001$
ABCEI**	$.855 \pm .001$	$.005 \pm .001$	$.849 \pm .001$	$.006 \pm .001$
ABCEI	$\mathbf{.871 \pm .001}$	$\mathbf{.003 \pm .001}$	$\mathbf{.863 \pm .001}$	$\mathbf{.005 \pm .001}$

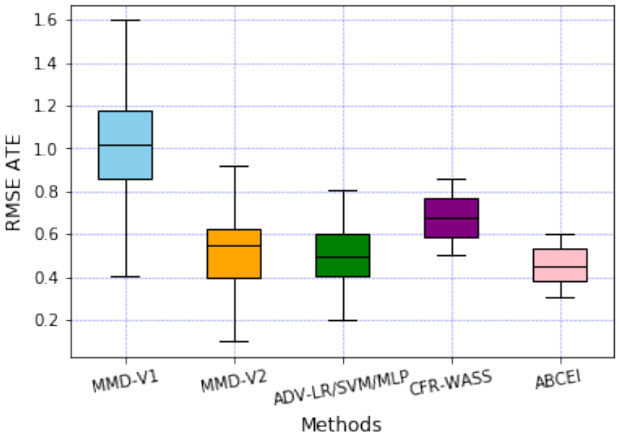


Figure 4.4: Results on ACIC datasets.

timator ensures that the encoder learns representations preserving useful information from the original covariate space. The adversarial learning component constrains the encoder to learn balanced representations. This causes ABCEI to achieve better performance than the baselines. We also report the performance of our model without mutual information estimator or adversarial learning, respectively, as ABCEI*, ABCEI**. From the results we can see that performance suffers when either of these components is left out, which demonstrates the importance of combining adversarial learning and mutual information estimation in ABCEI.

In Figure 4.4, we compare ABCEI with recent balancing methods on ACIC benchmark. As we can see, the variance of representation learning methods are lower than methods reweighing samples on covariate space. We also found that the adversarial balancing methods perform better on ATE estimation. ABCEI has the advantage of adversarial balancing as well as preserving predictive information in latent space, which makes it outperforms the other baselines.

4.6.5 Robustness Analysis on Selection Bias

To investigate the performance of our model when varying the level of selection bias, we generate toy datasets by varying the discrepancy between the treatment and control groups. We draw 8 000 samples with ten covariates $x \sim \mathcal{N}(\mu_0, 0.5 \cdot (\Sigma + \Sigma^T))$ as control group, where $\Sigma \sim \mathcal{U}((-1,1)^{10\times 10})$. Then we draw 2 000 samples from $x \sim \mathcal{N}(\mu_1, 0.5 \cdot (\Sigma + \Sigma^T))$. By adjusting μ_1,

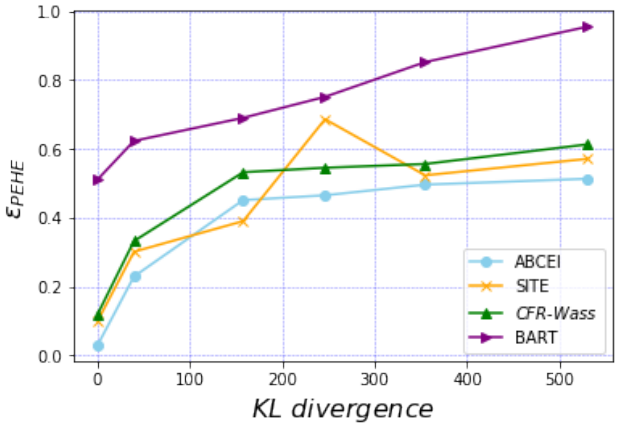


Figure 4.5: ϵ_{PEHE} on datasets with varying treatment selection bias. ABCEI is comparatively robust.

we generate treatment groups with varying selection bias, which can be measured by KL-divergence. For the outcomes, we generate $y|x \sim (w^T x + n)$, where $n \sim \mathcal{N}(0^{2\times 1}, 0.1 \cdot I^{2\times 2})$ and $w \sim \mathcal{U}((-1,1)^{10\times 2})$.

In Figure 4.5, we can see the robustness of ABCEI, in comparison with CFR-Wass, BART, and SITE. The reported experimental results are averaged over 100 test sets. From the figure, we can see that with increasing KL-divergence, our method achieves more stable performance. We do not visualize standard deviations as they are negligibly small.

4.6.6 Robustness Analysis on Mutual Information Estimation

To investigate the impact of minimizing the information loss on causal effect learning, we block the adversarial learning component and train our model on the IHDP dataset. We record the values of the estimated MI and ϵ_{PEHE} in each epoch. In Figure 4.6, we report the experimental results averaged over 1000 test sets. We can see that with increasing MI, the mean square error decreases and reaches a stable region. But without the adversarial balancing component, the ϵ_{PEHE} cannot be further lowered due to the selection bias. This result indicates that even though the estimators benefit from highly predictive information, they will still suffer if imbalance is ignored.

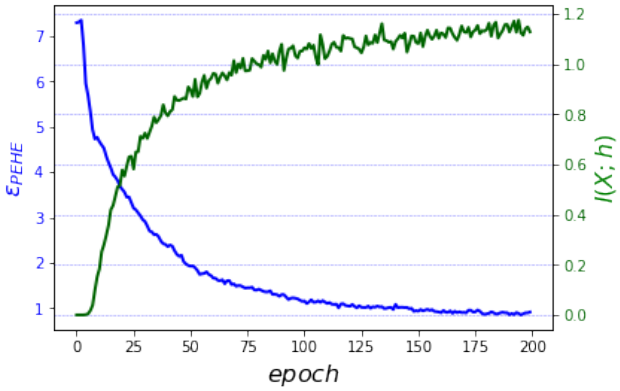


Figure 4.6: Mutual information (MI) between representations and original covariates, as well as ϵ_{PEHE} in each epoch. With increasing MI, ϵ_{PEHE} decreases.

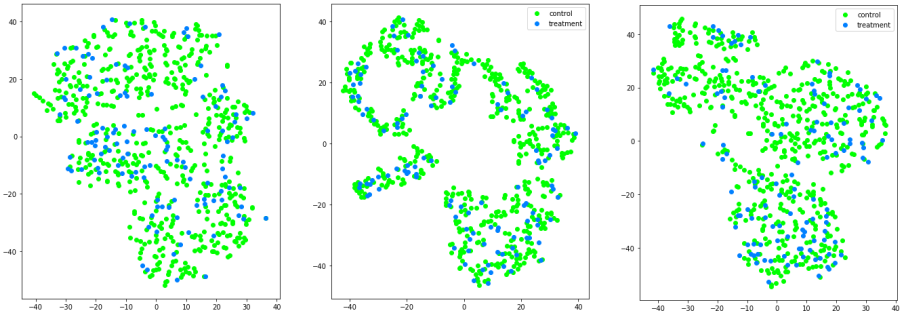
4.6.7 Balancing Performance of Adversarial Learning

In Figure 4.7, we visualize the learned representations on the IHDP and Jobs datasets using t-SNE. We can see that compared to CFR-Wass, the coverage of the treatment group over the control group in the representation space learned by our method is better. This showcases the degree to which adversarial balancing improves the performance of ABCEI, especially in population causal effect (ATE, ATT) inference.

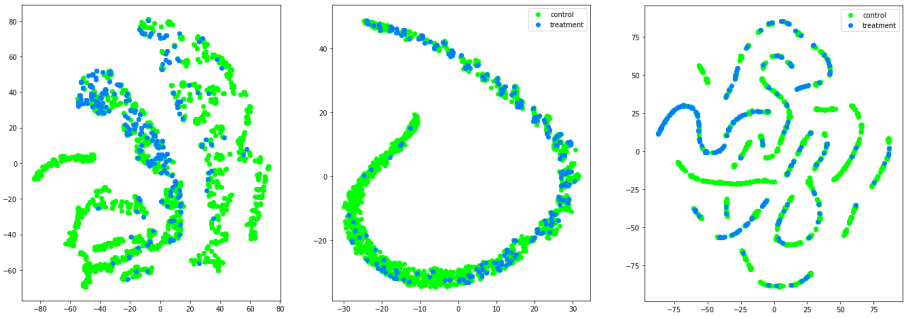
4.7 Conclusion

In this chapter, we study the modeling of causal dependency and how to capture the uncertainty in causal dependency modeling. This study could provide us a vision on the form of model class in EMM. For instance, we can employ a causal model in EMM to investigate whether a new drug would be exceptionally effective or ineffective on some subgroups.

To enable such an investigation, we need to develop tools for the estimation of causal effects. We propose a novel method for causal effect inference with observational data, called *ABCEI*, which is built on deep representation learning methods. ABCEI focuses on balancing latent representations from treatment and control groups by designing a two-player adversarial game. We use a discriminator to distinguish the representations from different groups. By adjusting the encoder parameters, our aim is to find an encoder that can fool



(a) IHDP.



(b) Jobs.

Figure 4.7: t-SNE visualization of treatment and control group, on the IHDP and Jobs datasets. The blue dots are treated units, and the green dots are control units. The left figures are the units in original covariate space, the middle figures are representations learned by ABCEI, and the right figures are representations learned by CFR-Wass; notice how the latter has control unit clusters unbalanced by treatment observations.

the discriminator, which ensures that the distributions of treatment and control representations are as similar as possible. Our balancing method does not make any assumption on the form of the treatment selection function. With the mutual information estimator, we preserve highly predictive information from the original covariate space to latent space. Experimental results on benchmark datasets and synthetic datasets demonstrate that ABCEI is able to achieve robust, and substantially better performance than the state of the art.

5

Uncertainty in Local Causal Dependency

"It is not that the meaning cannot be explained. But there are certain meanings that are lost forever the moment they are explained in words."

1Q84,
Haruki Murakami, 2009.

5.1 Introduction

In previous research, we assume that subgroups in terms of attribute variables are highly related with the exceptional performance of the models. This is the most important assumption for Exceptional Model Mining. Based on this assumption, we can start to construct the search space for EMM. Most of the previous research on EMM evaluate the contribution of each attribute variable and its domain by empirically measuring the qualities. Heuristic or exhaustive searching algorithms could lead us to the optimal solution. However, for the heuristic search process, some information could be lost; for the exhaustive search process, it is unfeasible to enumerate all the patterns on large scale datasets. Belfodil et al. (2018) propose anytime subgroup discovery to provide guarantees for bounding the errors of quality and show how far the quality can be from the best. However, the search algorithm is assumed to go over all the attribute features with the assumption that all the features are correlated to the target of interest. Knowing the relations between attributes and targets can help us model the dependencies between subgroups and the performance of the models. This dependency is called Local Causal Dependency.

Definition 5.1.1 (Local Causal Dependency) Assume we have a dataset $\Omega \sim P(X, Y, Z)$, where X, Y are sets of target variables, Z is a set of attribute variables. Model Φ is a mapping function $\Phi : X \rightarrow Y$. A subgroup S_D is

defined in terms of a description language with values taking from restricted domains $D(Z)$. The Local Causal Dependency is a stochastic process that determines the distribution of Y conditioning on X: $P\Big(Y|X, do\big(D(Z)\big)\Big)$,

where $do(\cdot)$ is an operator that does intervention on subgroup-level. Properly capturing the uncertainty in Local Causal Dependency can leverage EMM from the level of correlations to the level of causations. To realize such an updating, new model classes and quality measures for modeling and comparing Local Causal Dependency are needed. We introduce in this chapter how to estimate the quantity of Local Causal Dependency and how to measure the quality of subgroups with regard to Local Causal Dependency.

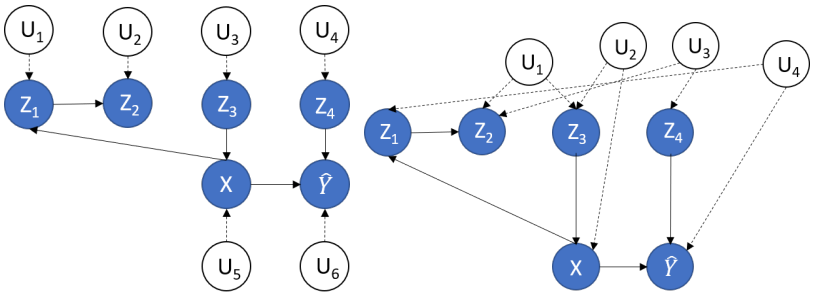
With the development of machine learning research, there is emergent requirement on the explaination of decision making process rather than just the performance of a model. In this chapter, we consider this problem as a local pattern mining task with EMM framework. In this task, multiple output variables depend on multiple input variables, and interestingness (model's performance is substantially different) is gauged in terms of some (to be instantiated) interactions between the output variables. We call such a dependency as Local Causal Dependency. Then we propose D-graph, a causal graph with extra nodes pointing to descriptive variables, which indicate the change of local mechanisms, charactering the dissimilarity of statistical quantity between subgroups as the dissimilarity of the associated causal models. We further propose to leverage functional constraints to compute Local Causal Dependency in the presence of unobserved confounders and to boost the subgroup search process, with respect to associated causal graph. To measure the difference of statistical quantities within and without subgroups, we propose an information-theoretic quality measure. Experiments on synthetic data show that our method outperforms the causality-oblivious baseline in terms of AUC, with an ROC curve that dominates the baseline ROC curve. Also, our method scales much better than the baseline in terms of both the number of attributes and the number of records: handling causality with care enables Exceptional Model Mining on larger datasets.

5.2 Motivation

With the rapid development of machine learning research, people start to focus on uncovering the black-box of decision making process rather than only the performance of models (Lakkaraju et al., 2017). Interpretable machine learn-

ing research can throughly boost and improve the fairness, accountability and transparency (Lepri et al., 2018) of models on real-world applications such as health care (Panigutti et al., 2019) and financial policy (Chen et al., 2018b). Most of these research try to find important features and establish a connection between those features and the outcome of model predictions (Doshi-Velez and Kim, 2017). Several techniques are developed to fulfill this task, e.g. Zintgraf et al. (2017) propose to explain the decision making process of deep neural networks by pixel-level salient regions in the input images; Ribeiro et al. (2016) propose to employ a surrogate model to find the relation between the rank of feature importance and the predictive outcome; instead of using low-level input features, Dash et al. (2018) propose to generate column based high-level features for the explain of binary classification. Different from those methods, we consider the decision making of a machine learning model from the view of local pattern mining (Morik et al., 2005). Properly finding how local patterns can influence the decision of a model is non-trivial, e.g. Kearns et al. (2017) points out that fairness manipulating by only considering the pre-defined subgroups may bring more biases to the decision making process. Well defined local pattern mining framework such as Subgroup Discovery (Atzmueller, 2015) and Exceptional Model Mining (Duivesteijn et al., 2016) provide us powerful tools to tackle this task, e.g. Duivesteijn and Thaele (2014) propose to understand where the classifier does (not) work and Grünwald and Grunwald (2007) propose to interpret the classification outcome by local patterns discovered with Minimum Description Length (MDL). Conceptually, Caruccio et al. (2015) formulates this task as Functional Dependencies (FD). However, there is one main disadvantage for those methods, that is, they only consider the correlation between features and the decision making process. Variables that are highly associated with each other do not mean that there are causal relations between them. The reason might be that they are confounded by the third part of variables. The state-of-the-art interpretable methods never solved this problem. Properly solving this problem could prevent us from being misled by spurious associations between features and the decision making process, which can leverage our research on explainable machine learning from the level of correlations to the level of causations.

In this chapter, we investigate the decision making process as a statistical quantity $P(\hat{Y}|X)$, e.g. a prediction model denoting a mapping from $\mathcal{X}$ to $\mathcal{Y}$. We assume that a third part of variables Z are highly associated with the decision making process. Subgroups are defined in terms of Z. Our aim is to discover interesting subgroups for which the decision making process (the statistical quantity $P(\hat{Y}|X)$) is substantially different from the decision making



(a) Causal graph $\mathcal{G}_0$ without hidden confounders. **(b)** Causal graph $\mathcal{G}$ with hidden confounders.

Figure 5.1: Causal graphs.

process on the whole dataset. This task is a standard subgroup discovery / exceptional model mining task. Previous research for this problem is oblivious to the structural relation between variables, which could lead to the following problems: 1) spurious association may mislead the algorithm to discover subgroups / patterns with false interestingness; 2) the search space indicated by spurious association would be very large. In contrast with previous research, we consider the *causal* mechanism between the third part of variables Z and the statistical quantity $P(\hat{Y}|X)$. In particular, we make contribution to current state-of-the art in two hands: 1) on the one hand, the causal dependency we investigate is a little different from classical causal effects between two (groups of) variables. Rather than causal quantity $P(Y|do(X = x))$, we consider an intervention on the third part of variables Z and the effects on a statistical quantity $P(\hat{Y}|X)$. We call such kind of dependencies as *local causal dependencies* on the subgroup-level. 2) On the other hand, we consider such *local causal dependencies* in the presence of unobserved confounders. This would bring more challenges for the estimate of causal quantity, e.g. in Figure 5.1a, we show a causal graph $\mathcal{G}_0$ without hidden confounders: the randomness is provided by independent unknown noises. In Figure 5.1b, the presence of unobserved variables may bring extra constraints to the causal quantities which cannot be captured by conditional independence. Hence, methods that can tackle the influence of unobserved confounders are required.

5.3 Contributions

In this chapter, we define a new intervention paradigm on subgroup-level in terms of a third part of variables Z and then estimate the differences between statistical quantities $P(\hat{Y}|X)$ by the differences between causal graphs in subgroups and in the whole dataset. The model class and quality measures proposed in previous research of EMM never solved this problem. Our method can provide more insights on why and how local patterns influence the performance of a model. The main contributions are:

- We leverage the underlying causal mechanism to model the decision making process. This can boost local pattern mining methods like EMM, preventing the algorithm from being misled by spurious associations.
- We define a new intervention paradigm on the subgroups, which can explain how causal dependencies between the third part of attribute variables Z influence the decision model $P(\hat{Y}|X)$. We call this relation *local causal dependencies*.
- We consider computing the *local causal dependencies* in the presence of unobserved confounders, which can help us refine the beam search algorithms for finding most interesting subgroups. This tackles the problem that is never solved by previous research on local pattern mining.
- Experimental results on both synthetic and real-world datasets demonstrate the effectiveness of our method quantitatively and qualitatively.

5.4 Related Work

The natural property of interesting descriptions draw a connection from LPM to interpretable machine learning. Some focus on investigating the disadvantage subgroups to analyze the fairness of network representation model considering the local structures (Du et al., 2020c); Some focus on finding reliable functional dependencies between variables (Mandros et al., 2017). However, the main drawback of these methods is that they might be misled by the spurious associations between variables. We tackle this problem by providing a new intervention paradigm on local patterns which can adapt causal dependencies to subgroup-level.

Traditional model classes in EMM (Duivesteijn et al., 2010, Lemmerich et al., 2016) propose to measure the difference between graphs like Bayesian networks. However the measure is observational equivalent, which cannot distinguish the difference derived from causal graphs. Unobserved confounders

in the causal graph could bring uncertainty to the estimate of causal quantities. New constraints associated with unobserved confounders cannot be captured by conditional independence (Verma and Pearl, 1991). Some research proposes to leverage c-components decomposition (Tian and Pearl, 2002b) to capture such constraints; some focus on building extra variables to model such constraints (Chen et al., 2017). In this chapter, we look for constraints in the presence of unobserved confounders that are related with the *local causal dependencies*, which can help us prevent the search algorithms from being misled by spurious associations.

For the interestingness measure, domain experts may want to learn how these patterns changing across different groups could help them to understand why their classifiers perform differently with an interpretable answer. This can also help to understand fairness in machine learning models (e.g. classifiers) across different subgroups (Choi et al., 2019). Traditional quality measures like WRAcc (van Leeuwen and Knobbe, 2011), z-score (van Leeuwen and Knobbe, 2012), and KL-divergence (Mampaey et al., 2015) cannot be qualified to measure the differences between the *local causal dependencies*. The quality measure used in this chapter is built on information theory (Janzing et al., 2019) considering the mean distance of feature vectors in Reproducing Kernel Hilbert Space (RKHS) (Smola et al., 2007). This allows us to measure the difference between conditional distributions using Integral Probability Metric (IPM) (Sriperumbudur et al., 2010). By considering the autonomous mechanism of causal structure, we show shat the independence relation appear in the quality scores, following (Janzing et al., 2019). On the other hand, we leverage functional constraints to decompose the quantity of interest, so that we can compute the statistical quantity by only reweighing the quantity with constraints that are changed in the subgroups. This can throughly improve the running speed of our algorithm.

5.5 Methodology

5.5.1 Preliminaries

Assume a set of descriptive variables $Z = \{z_1, \cdots, z_k\}$ and two sets of target variables $X = \{x_1, \cdots, x_\ell\}$, $Y = \{y_1, \cdots, y_m\}$. The observational dataset Ω is drawn from a distribution $P(\Omega)$. For a given sample Ω, we have $P(V) = P(Z, X, Y)$ consisting of a bag of N records $r^i = (Z^i, X^i, Y^i)$. By assuming that values of Z are taken from an unrestricted domain $\mathcal{A}$, we can define a

function $D : \mathcal{A} \rightarrow \{0, 1\}$. A description D covers a record r^i if and only if $D(z_1^i, \cdots, z_k^i) = 1$.

Structural Causal Model We use the Structural Causal Model (SCM) (Pearl, 2009) as the tool to model the decision making process regarding to the underlying causal mechanism that determines the distribution $P(Z, X, \hat{Y})$:

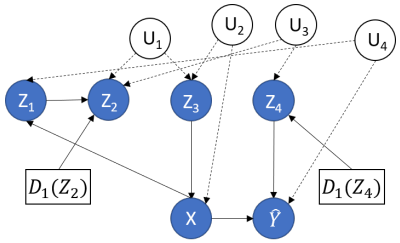
Definition 5.5.1 (Structural Causal Model) (Pearl, 2009) A structural causal model M is represented by a 4-tuple $\langle V, U, \mathcal{F}, P(U)\rangle$ where:

1. U is a set of exogenous (unobserved) variables of any types including continuous, discrete, or mixed;
2. V is a set of endogenous (observed) variables;
3. $\mathcal{F}$ is a set of functions $\mathcal{F} = \{f_i\}$ mapping from $V \cup U$ to V. For each endogenous variable $V_i \in V$, there is a function $f_i \in \mathcal{F}$ mapping from $Pa_i \cup U_i$ to V_i, where $Pa_i \subseteq (V \backslash V_i)$ stands for direct parents of V_i in the causal graph, and $U_i \subseteq U$ stands for sources of randomness that determine V_i;
4. $P(U)$ is a joint distribution over exogenous variables U, encoding the randomness.

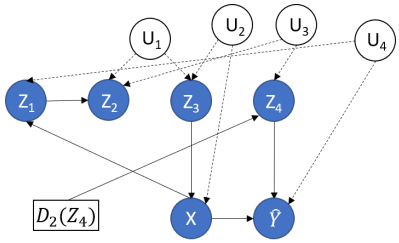
In this chapter, we do not make any assumption about the functional type of each $f_i \in \mathcal{F}$. We consider non-parametric causal relations between variables. A causal model M is associated with a causal graph $\mathcal{G}$ over the set of nodes V and U. We define $Pa(X)_{\mathcal{G}}$, $Ch(X)_{\mathcal{G}}$, $An(X)_{\mathcal{G}}$, $De(X)_{\mathcal{G}}$ as the union in $\mathcal{G}$ of $X \subseteq V$ with their parents, children, ancestors, and descendants, respectively. Each directed edge represents dependencies between variables and their parents, quantifying the conditional probabilities $P(v_i|pa_i, u_i)$, which implies an important property for SCM: the local autonomous mechanism (Peters et al., 2017). This property allows us to decompose the joint distribution $P(V)$ into:

$$P(V) = \sum_{u} \prod_{\{i|V_i \in V\}} P(v_i|pa_{v_i}, u_i) \prod_{\{i|U_i \in U\}} P(u_i|pa_{u_i}),$$

where the summation considers all the possibilities of unobserved variables. If there are no unobserved confounders for each observed node $V_i \in V$, the causal model satisfies the Markovian property (cf. Figure 5.1a). While real applications might include unobserved confounders (cf. Figure 5.1b), which is why we call these types of model *non-Markovian causal models*.



(a) D-graph $\mathcal{G}^*_{D_1}$ associated with D_1 : $d(z_2) = \{1, 2, 3\} \wedge d(z_4) = \{2, 3\}$.



(b) D-graph $\mathcal{G}^*_{D_2}$ associated with D_2 : $d(z_4) = \{2, 3\}$.

Figure 5.2: Description-enhanced causal graphs.

5.5.2 Why Does the Model Perform Differently?

We assume that according to the domain knowledge, for each given dataset $\Omega \sim P(X, Y, Z)$, we have a causal graph $\mathcal{G}$ that depicts the underlying decision making process (w.r.t prediction mapping Φ) with causal model M, generating the distribution $P(X, \hat{Y}, Z)$. For each subgroup S_D defined in terms of attributable variable Z, we can define a potentially different causal model M_D represented by graph $\mathcal{G}^*_D$, where extra edges $D(Z_i)$ represent that the local mechanism in the subgroup is different with the whole data. The links between D nodes and the observed nodes in $\mathcal{G}^*_D$ represent additional restrictions on the distributions of variables. We call this type of graph *D-graph*. Formally, we have the following definition:

Definition 5.5.2 (D-graph) For any description $D = \wedge_{\{i|z_i \in Z\}} d(z_i)$, $\forall z_i \in Ch(D)$, we have

$$p^*(z_i|pa_{z_i}, u_{z_i}) = \begin{cases} p_{d(z_i)}(z_i|pa_{z_i}, u_{z_i}) & \text{if } z_i \in d(z_i), \\ 0 & \text{if } z_i \notin d(z_i), \end{cases}$$

where $d(z_i)$ denotes the restricted domains of z_i, if $d(z_i) = \mathcal{A}$; and $p_{d(z_i)}$ represents the renormalized distribution regarding to values in the associated domain. Under this definition, the joint distribution $P_{M_D}(V)$ from the causal model M_D associated with the D-graph $\mathcal{G}^*_D$ can be decomposed by replacing

restricted components:

$$P_{M_D}(V) = \sum_{\mathbf{u}} \prod_{\{i|V_i \notin Ch(D)\}} p(v_i|pa_{v_i}, u_i) \prod_{\{j|V_j \in Ch(D)\}} p^*(v_j|pa_{v_j}, u_j)\, p(\mathbf{u}), \tag{5.1}$$

This decomposition implies that by substituting the equations in M, The D nodes modify the original causal model denoted by M_D. In Figure 5.2 we give two examples for D-graphs, where local patterns representing by D nodes point to associated observed nodes in the original graph. In Figure 5.2a, the domains of z_2 and z_4 are restricted to $\{1, 2, 3\}$ and $\{2, 3\}$. Domains of other attributable variables are kept unrestricted. For instance, $\mathcal{G}^*_{D_1}$ in Figure 5.2a indicates that $P_{M_{D_1}}(z_2|z_1) \neq P_M(z_2|z_1)$ and $P_{M_{D_1}}(z_4) \neq P_M(z_4)$. In order to know how interesting the differences between quantity of interest are, we need to define a quality measure, e.g. a function $\varphi\big(P_M(\hat{Y}|X) \,||\, P_{M_D}(\hat{Y}|X)\big)$ The EMM considering Local Causal Dependency can be reformulated as:

Problem 5.5.1 Given a dataset Ω, a mapping function $\Phi : \mathcal{X} \to \mathcal{Y}$, a quality measure function $\varphi(\cdot||\cdot)$, a causal graph $\mathcal{G}$ and its associated causal model M, we aim to find a sequence of Q descriptions $h = \{D_1, \cdots, D_Q\}$, such that $\forall D' \in \mathcal{D} \setminus h, \varphi\big(P_M(\hat{Y}|X) \,||\, P_{M_{D'}}(\hat{Y}|X)\big) < \varphi\big(P_M(\hat{Y}|X) \,||\, P_{M_D}(\hat{Y}|X)\big)$, $\forall D \in h$.

By defining D-graph, we build a connection between local pattern and the quantity of interest in causal graph language. We call it Local Causal Dependency. The D-nodes indicate why and how quantity of interest might be different within and without subgroups (cf. equation 5.1).

Local Causal Dependency in Graph Language Example 1 shows a problem that tradition EMM would have by ignoring the underlying mechanism. In this section, we introduce how to tack this problem by graph language. We assume the causal graph $\mathcal{G}$ associating with data generating mechanism in Figure 5.3. From $\mathcal{G}$ we can see that variable Z_3 can influence $\hat{Y}$ through X, but it is not a direct cause of $\hat{Y}$. This indicates that there is correlation between Z_3 and quantity $P(\hat{Y}|X)$ from observation, but if we do intervention on Z_3, the quantity will not change. We explain why we reach this conclusion.

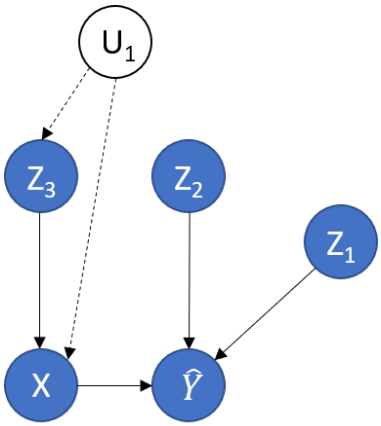


Figure 5.3: Causal graphs $\mathcal{G}$ of National Supported Work program.

Definition 5.5.3 (Equivalent Description) Two descriptions D_i and D_j are called *equivalent* relative to $P(\hat{Y}|X)$ if $Ch(D_i) \subset Ch(D_j)$ and $P_{M_{D_i}}(\hat{Y}|X) = P_{M_{D_j}}(\hat{Y}|X)$, with regard to D-graphs $\mathcal{G}^*_{D_i}$ and $\mathcal{G}^*_{D_j}$.

Definition 5.5.4 (Minimal Description) A description D is called *minimal* relative to $P(\hat{Y}|X)$ if there is no $Ch(D') \subset Ch(D)$ such that for each causal model M associated with causal graph $\mathcal{G}$, $P_{M_{D'}}(\hat{Y}|X) = P_{M_D}(\hat{Y}|X)$.

Definitions 5.5.3 and 5.5.4 jointly provide two levels of meanings: on the one hand, we can simplify a given description by looking for its minimality. This can prevent us from redundantly refining the descriptions, e.g., exploring the attributable spaces which are independent with the quantity of interests. On the other hand, we can measure the interestingness of descriptions considering not only the computed quantity of interests, but also structural relations retrieved from the causal graph.

One common method to generate a minimal description is using d-separation, conditioning on X, looking for variables that are independent with $\hat{Y}$. In the presence of unobserved confounders, some constraints cannot captured only by conditional independence (Verma and Pearl, 1991). Some research (Tian and Pearl, 2002b, Chen et al., 2017) propose systematic methods to find such constraints, functional constraint. In this chapter, we employ functional constraints to help us find minimal descriptions. In order to systemically find minimal descriptions, we need to recursively partition the observed variables (V) into groups by applying the confounded-component (c-component)

decomposition (Tian and Pearl, 2002a):

$$P(V) = \prod_j \sum_{\mathbf{u_j}} \prod_{\{i|V_i \in C_j\}} P(v_i|pa_i, u_i) P(\mathbf{u_j}).$$

By this definition, in the causal graph $\mathcal{G}_{\underline{V}}$, any two nodes in the same c-component may be confounded by unobservables. For instance, in Figure 5.3, the nodes can be partitioned by c-components as $\{Z_3, X\}$, $\{Z_1\}$, $\{Z_2\}$, $\{\hat{Y}\}$. Without loss of generality, we start from this example to show how to generate functional constraints for the quantity of interest using c-component decomposition. At first we can derive the joint distribution of $\hat{Y}, X$ as:

$$\begin{aligned}
P(\hat{Y}, X) &= \sum_{z_1,z_2,z_3} P(V) = \sum_{z_1,z_2,z_3} R(Z_3, X) R(Z_1) R(Z_2) R(\hat{Y}) \\
&= \sum_{z_1,z_2,z_3} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1} P(x, z_3|u_1) P(u_1) \\
&= \sum_{z_1,z_2,z_3} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1} P(x|z_3, u_1) P(z_3|u_1) P(u_1) \\
&= \sum_{z_1,z_2} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1,z_3} P(x|z_3, u_1) P(z_3|u_1) P(u_1),
\end{aligned}$$

such that the quantity of interest can be computed as:

$$\begin{aligned}
&P(\hat{Y}|X) = \\
&\frac{\sum_{z_1,z_2} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1,z_3} P(x|z_3, u_1) P(z_3|u_1) P(u_1)}{\sum_{z_1,z_2,z_3,y} P(V)} = \\
&\frac{\sum_{z_1,z_2} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1,z_3} P(x|z_3, u_1) P(z_3|u_1) P(u_1)}{\sum_{z_1,z_2,\hat{y}} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2) \sum_{u_1,z_3} P(x|z_3, u_1) P(z_3|u_1) P(u_1)} = \\
&\sum_{z_1,z_2} P(\hat{y}|x, z_1, z_2) P(z_1) P(z_2),
\end{aligned}$$

which implies that $P(\hat{Y}|X)$ is the functional of Z_1, Z_2. This functional constraint is consistent with the assumed model in Example 1. Here, two main properties of c-component decomposition are applied. First, we can decompose the joint distribution into product of conditional distributions of each c-component; on the other hand, each c-component is only dependent on its non-descendant variables in the c-component and effective parents of its non-descendant variables in the c-component (Tian and Pearl, 2002b). Inspired by those properties, we can derive the following theorem for computing the quantity of interest with c-components:

Theorem 5.5.1 (functional constraint set (FCS)) Let $X, \hat{Y} \subseteq V$ be disjoint sets of variables. Let $W = An(X \cup \hat{Y})_{\mathcal{G}}$ be partitioned into c-components $C(W) = \{C_1(W_1), \cdots, C_J(W_J)\}$ in causal graph $\mathcal{G}_{[An(X \cup \hat{Y})]}$. Then the quantity of interest $P(\hat{Y}|X)$ is the function of $W' = \bigcup_h C_h(W_h) \setminus X \subseteq W \setminus X$, if and only if

$$\forall C_h(W_h) \in C(W), An^+(\hat{Y})_{G_{[An(X \cup \hat{Y})]\underline{X}}} \cap C_h(W_h) \neq \emptyset,$$

where $An^+(Y)_{G_{[An(X \cup \hat{Y})]\underline{X}}}$ represents the ancestor set of $\hat{Y}$ including $\hat{Y}$, in graph $G_{[An(X \cup \hat{Y})]\underline{X}}$, the subgraph $\mathcal{G}_{[An(X \cup \hat{Y})]}$ removing all out edges relative to X.

Proof. According to Bayes equation, we have $P(Y|X) = \frac{P(X,Y)}{P(X)}$. By c-component decomposition (Tian and Pearl, 2002a), $P(X, Y)$ can be decomposed into:

$$\sum_{z \setminus z'} R(X \setminus X', Z \setminus Z') R(Y) \sum_{z'} R(X'). \tag{5.2}$$

$P(X)$ can be decomposed into

$$\sum_{y, z \setminus z'} R(X \setminus X', Z \setminus Z') R(Y) \sum_{z'} R(X'), \tag{5.3}$$

where $R(X \setminus X', Z \setminus Z')$ share nodes with $An^+(Y)_{G_{[An(X \cup \hat{Y})]\underline{X}}}$. According to (Tian and Pearl, 2002b, Lemma 2), $\sum_{z'} R(X')$ is not a function of Y, so that it can be removed by divide operator. Hence we have $P(Y|X) = \frac{\sum_{z \setminus z'} R(X \setminus X', Z \setminus Z') R(Y)}{\sum_{y, z \setminus z'} R(X \setminus X', Z \setminus Z')}$, such that $z \setminus z'$ denote the functional constraints, which are the intersections of c-components with Ancestor in the causal graph $G_{[An(X \cup Y)]\underline{X}}$. ■

Theorem 5.5.1 implies that even if a variable is not the ancestor of prediction variable $\hat{Y}$, it can also affect the quantity of interest $P(\hat{Y}|X)$. In the following, we give a more general example to show this property:

Example 2 Figure 5.4 shows a causal graph $\mathcal{G}$ and its disconnected subgraphs in $\mathcal{G}_{\underline{V}}$, from which we can have the c-components $\{X_1, Z_1, Z_2\}$, $\{\hat{Y}\}$ and $\{X_2, Z_3\}$ (cf. Figure 5.4b). By c-components decomposition, we have:

$$P(X,\hat{Y}) = \sum_{z_1,z_2,z_3} R(X_1,Z_1,Z_2)R(\hat{Y})R(X_2,Z_3)$$
$$= \sum_{z_1,z_2} R(X_1,Z_1,Z_2)R(\hat{Y}) \sum_{z_3} R(X_2,Z_3),$$

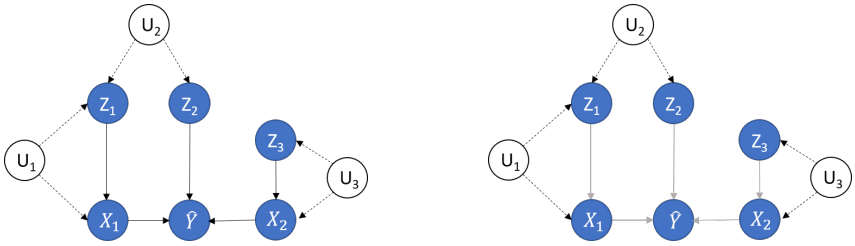
$$P(X) = \sum_{z_1,z_2,z_3,\hat{y}} R(X_1,Z_1,Z_2)R(\hat{Y})R(X_2,Z_3)$$
$$= \sum_{z_1,z_2,\hat{y}} R(\hat{Y})R(X_1,Z_1,Z_2) \sum_{z_3} R(X_2,Z_3),$$

$$\begin{aligned} P(\hat{Y}|X) &= \frac{\sum_{z_1,z_2} P(\hat{y}|z_2,x_1,x_2)R(X_1,Z_1,Z_2)}{\sum_{z_1,z_2,\hat{y}} P(\hat{y}|z_2,x_1,x_2)R(X_1,Z_1,Z_2)} \\ &= \frac{\sum_{z_1,z_2} P(\hat{y}|z_2,x_1,x_2)R(X_1,Z_1,Z_2)}{\sum_{z_1,u_1} P(x_1|u_1,z_1)P(z_1|u_1)P(u_1)} \\ &= \frac{\sum_{z_1,z_2} P(\hat{y}|z_2,x_1,x_2)R(X_1,Z_1,Z_2)}{\sum_{z_1} P(x_1|z_1)P(z_1)} \\ &= \frac{\sum_{z_1,z_2} P(\hat{y}|z_2,x_1,x_2)P(x_1|z_1,z_2)P(z_1|z_2)P(z_2)}{\sum_{z_1} P(x_1|z_1)P(z_1)} \\ &= \sum_{z_2} P(\hat{y}|z_2,x_1,x_2)P(z_2)\frac{\sum_{z_1} P(x_1|z_1)P(z_1|z_2)}{\sum_{z_1} P(x_1|z_1)P(z_1)} \end{aligned} \tag{5.4}$$

where

$$R(X_2,Z_3) = \sum_{u_3} P(x_2|z_3,u_3)P(z_3|u_3)P(u_3),$$
$$R(X_1,Z_1,Z_2) = \sum_{u_1,u_2} P(x_1|u_1,z_1)P(z_1|u_1,u_2)P(z_2|u_2)P(u_1)P(u_2).$$

The presence of hidden confounder indicates $Z_1 \not\perp\!\!\!\perp Z_2|U_2$, which would not allow Equation (5.4) to be reduced further. In Figure 5.4a, we have $An^+(\hat{Y})_{G_{[An(X\cup\hat{Y})]\underline{X}}} = \{Z_2,\hat{Y}\}$. According to Theorem 5.5.1, $\{X_1,Z_1,Z_2\}$ is the only c-component to construct functional constraint set, because $\{X_1,Z_1,Z_2\} \cap \{Z_2,\hat{Y}\} = \{Z_2\}$. This implies that the quantity of interest is the function of Z_1 and Z_2, which is consistent with the results in Equation (5.4). Now we can derive an important property for minimal description:



(a) $\mathcal{G}$ **(b)** $\mathcal{G}_{\underline{V}}$

Figure 5.4: Causal Graphs.

Corollary 5.5.1 (Minimality) D is a minimal description satisfying minimality for D-graph $\mathcal{G}_D^*$ if and only if $Ch(D) \subseteq W'$.

Corollary 5.5.1 implies that we can validate whether a description is a minimal description only by graph criterion, without computing the quantity of interest for each candidate subgroup.

5.5.3 Information Theoretic Quality Measure

The quality measure in this chapter is proposed based on measuring the differences between quantity of interest $P_M(\hat{Y}|X)$ in the whole data and $P_{M_D}(\hat{Y}|X)$ in the subgroup in terms of description D. Considering the complexity of $P(\hat{Y}|X)$, we propose to use Integral Probability Metric (IPM) (Müller, 1997) to quantify the dissimilarity between conditional distributions:

$$\varphi_\Theta(\mathbb{P}||\mathbb{Q}) = \sup_{\vartheta \in \Theta} \Big| \mathbb{E}_\mathbb{P}[\vartheta|X] - \mathbb{E}_\mathbb{Q}[\vartheta|X] \Big|,$$

where $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$, $\mathcal{P}$ is the set of all Borel probability measures on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A})$, and Θ is a class of bounded real-valued measurable functions on $\mathcal{Y}$. Following (Gretton et al., 2007, Smola et al., 2007, Sriperumbudur et al., 2010), we choose to use Θ in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with k as reproducing kernel, such that:

$$\varphi_\Theta(\mathbb{P}||\mathbb{Q}) = \Big|\Big| \sum_{\hat{y},x} k(\cdot, \hat{y}) P_M(\hat{y}|x) - \sum_{\hat{y},x} k(\cdot, \hat{y}) P_{M_D}(\hat{y}|x) \Big|\Big|_\mathcal{H},$$

where $k(\cdot, \hat{y}) = \phi(\hat{y})$ is a feature map from $\mathcal{Y}$ to $\mathcal{H}$. Now we can quantify the exceptionality of the target quantities in subgroups using this distance measure

on probability distributions. A quality measure based on information theoretic exceptionalities can be defined as:

Definition 5.5.5 (Information Theoretic Exceptionality) Let Y be a random variable in dataset Ω, we can have conditional distribution of the predictive variable $\hat{Y}$, $P_M(\hat{Y}|X)$ with regard to a decision model M. Assume Ω is sampled from $P(\Omega)$, then we can have another conditional distribution $P'_M(\hat{Y}|X)$ sampled from $P(P_M(\hat{Y}|X))$. For a given subgroup in terms of description D, we have a distribution $P_{M_D}(\hat{Y}|X)$. We can define a distance measure $\varphi : \mathcal{X} \times \mathcal{Y} \times \mathcal{D} \to \mathbb{R}_0^+$, such that:

$$P\{\varphi\big(P_M(\hat{Y}|X)||P'_M(\hat{Y}|X)\big) \geq \varphi\big(P_M(\hat{Y}|X)||P_{M_D}(\hat{Y}|X)\big)\} = e^{-\varphi\left(P_M(\hat{Y}|X)||P_{M_D}(\hat{Y}|X)\right)},$$

where φ needs to be a surjective function. For simplicity of representation, we let $\varphi\big(P_M(\hat{Y}|X)||P_{M_D}(\hat{Y}|X)\big) = \varphi_{\hat{Y}|X}(D)$ and $\varphi\big(P_M(\hat{Y}|X)||P'_M(\hat{Y}|X)\big) = \varphi_{\hat{Y}|X}(\Omega)$.

This definition follows the formulation of hypothesis testing. By the definition, $P_M(\hat{Y}|X)$ and $P'_M(\hat{Y}|X)$ are drawn from same distribution, the cumulative probability $P\{\varphi_{\hat{Y}|X}(\Omega) \leq \varphi_{\hat{Y}|X}(D)\} = 1 - e^{-\varphi_{\hat{Y}|X}(D)}$ would be close to 1, if $\varphi_{\hat{Y}|X}(D)$ is extremely higher than expected, which implies that we can reject the hypothesis "$P^D_{\hat{Y}|X}$ and $P_{\hat{Y}|X}$ are drawn from the same distribution". The distance measure can be represented as:

$$\varphi_{\hat{Y}|X}(D) = -\log P\{\varphi_{\hat{Y}|X}(\Omega) \geq \varphi_{\hat{Y}|X}(D)\},$$

where D and Ω share the same support in the space of probability distribution $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Specifically, we can define $\varphi_{\hat{Y}|X}(D) = -\log P(P^D_{\hat{Y}|X})$, such that we have:

$$\varphi_{\hat{Y}|X}(D) = -\log P\{P(P_{\hat{Y}|X}) \leq P(P^D_{\hat{Y}|X})\}.$$

Here we define the exceptionality in terms of conditional distribution, following (Janzing et al., 2019), we have the following Lemma:

Lemma 5.5.1 (Exceptionality Independence) If $\varphi_{\hat{Y}|X=x}$ is a surjective Information Theoretic Exceptionality score with regard to conditional distribution $P_{\hat{Y}|X}$, then $\varphi_{\hat{Y}|X} \perp\!\!\!\perp X$.

This lemma implies that for all $x \in \mathcal{X}$, $\varphi_{\hat{Y}|X=x}$ has the same density. Now we can introduce how to leverage functional constraints from the causal model to compute quality measure for subgroups. Theorem 5.5.1, and Corollary 5.5.1 imply that each quantity of interest $P(\hat{Y}|X = x_i)$ can be represented as the form:

$$P(\hat{Y}|X) = \prod_{\{y_i \in Y\}} \sum_{\mathbf{z},\mathbf{z}'} P(y_i|x_i, \mathbf{z})P(\mathbf{z})Q(\mathbf{z}'), \tag{5.5}$$

where $\mathbf{z}$ represents a set of variables which are parents of $\hat{Y}_i$, $\mathbf{z}'$ represents variables which are parents of X, and there exists a hidden confounder between them and ancestors of $\hat{Y}_i$. $P(\mathbf{z})$ can be represented as the products of distributions $\prod_j P(z_j)$, and $Q(\mathbf{z}')$ can be represented as the form of $\frac{\sum_{\mathbf{z}'} P(X_i|\mathbf{z}')P(\mathbf{z}'|\mathbf{z})}{\sum_{\mathbf{z}'} P(X_i|\mathbf{z}')P(\mathbf{z}')}$. For each dataset $\Omega = (X, Y, Z)_n$ and associated graph $\mathcal{G}$, we can learn a mapping function $P(\hat{Y}|X, pa_{\hat{Y}})$ and compute $P_M(\hat{Y}|X)$ by averaging over $P(Z)$. For each given subgroup D, we can compute $P_{M_D}(\hat{Y}|X)$ by replacing $z \in Ch(D)_{\mathcal{G}^*_D}$ with $P^*(z)$, with regard to associated D-graph. Equation (5.5) also implies that, whenever there is a D-node pointing to z', $P(z'|z) = P(z')$, because the local mechanism for generating z' has changed. Hence, we can always compute $P(\hat{Y}|X)$ by using $P^*(\mathbf{z})$. For evaluation, we can draw m samples from $P_M(\hat{Y}|X = x)$ and n samples from $P_{M_D}(\hat{Y}|X)$, such that we can empirically estimate $\varphi_{Y|X}(D)$ using maximum mean discrepancy (MMD), inspired by (Gretton et al., 2012):

$$\begin{aligned}
&MMD_u^2[\Theta, P_M(\hat{Y}|X), P_{M_D}(\hat{Y}|X)] = \\
&\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \vartheta(\hat{y}_i, \hat{y}_j) + \\
&\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \vartheta(\hat{y}_i^D, \hat{y}_j^D) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \vartheta(\hat{y}_i, \hat{y}_j^D),
\end{aligned}$$

which is a sum of two U-statistics and a sample average. Note that for each dataset we feed $\{x\}$ in the whole dataset and subgroup, respectively, to get the predicted results $\{\hat{y}\}$ and $\{\hat{y}^D\}$.

By jointly applying Theorem 5.5.1, Corollary 5.5.1, we propose an algorithm 3 extended from (Duivesteijn et al., 2016, Algorithm 1) for finding top-Q exceptional subgroups.

Algorithm 3 Causality-aware beam search for top-Q exceptional model mining.

Input: Dataset Ω, Graph $\mathcal{G}$, Quality Measure φ, Refinement Operator η, Integer w, d, Q, c-components decomposition operator ψ
Output: PriorityQueue resultSet

function SABS(Ω, φ, η, w,d, Q, W'=FCS(Ω, $\mathcal{G}$, ψ))
    candidateQueue $\leftarrow$ new Queue;
    candidateQueue.enqueue({});
    resultSet $\leftarrow$ new PriorityQueue(Q);
    while level $\leq$ d **do**
        beam $\leftarrow$ new PriorityQueue(w);
        while candidateQueue $\neq \varnothing$ **do**
            seed $\leftarrow$ candidateQueue.dequeue();
            set $\leftarrow \eta$(seed);
            for all D $\in$ set **do**
                if $Ch(D)_{\mathcal{G}_D^*} \subseteq W'$ **then**
                    quality $\leftarrow \varphi(Ch(D)_{\mathcal{G}_D^*})$;
                    resultSet.insert_with_priority(D, quality);
                    beam.insert_with_priority(D, quality);
                end if
            end for
        end while
        while beam $\neq \varnothing$ **do**
            candidateQueue.enqueue(beam.get_from_element());
        end while
    end while
    return resultSet;
end function

function FCS($\Omega, \mathcal{G}$, ψ)
    $W' \leftarrow \{\}$;
    $W \leftarrow An(X \cup \hat{Y})_{\mathcal{G}}$;
    $C_1(W_1), \cdots, C_J(W_J) \leftarrow \psi(G_{[An(X \cup \hat{Y})]})$;
    for $h = 1$ to J **do**
        if $C_h(W_h) \cap An^+(\hat{Y})_{G_{[An(X \cup \hat{Y})]_{\underline{X}}}} \neq \emptyset$ **then**
            $W' \leftarrow W' \cup C_h(W_h) \setminus X$;
        end if
    end for
    return W';
end function

5.6 Experiments

In this section, we design various experiments in order to validate our method against the following questions:

RQ1 Comparing to quality measures that ignore the structural relations between variables, can our method reliably find the injected exceptional subgroup in synthetic dataset?

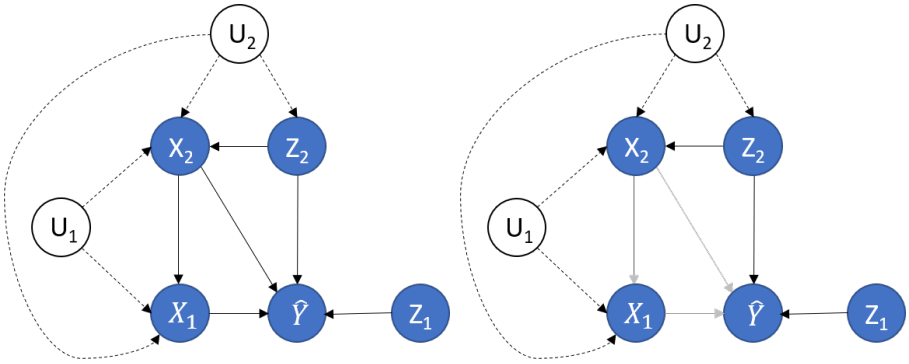
RQ2 Comparing to search algorithms that ignore the structural relations between variables, can our method improve the time efficiency?

RQ3 For real-world datasets in which the ground truth is unknown, can our algorithm effectively discovery exceptional subgroups?

Synthetic Dataset For the synthetic dataset, we propose to generate the data by the following steps: 1) we initialize nodes $\{x_1, \ldots, x_\ell\}$, $\{y_1, \ldots, y_m\}$, $\{z_1, \ldots, z_k\}$ with control parameters ℓ, m, k. For each node, the number i of parents is sampled with probability decaying inverse proportional to i. For each y, x, z, we sample parents from X and Z. For each pair of variables in $(x, z), (z, z)$ and (y, z) we sample a hidden confounder following Bernoulli distribution with parameter α. By these steps, we can randomly draw causal graphs for which the quantity $P_{M_D}(\hat{Y}|X)$ can always be computable with D-graph. 2) Given a sampled causal graph, we propose to generate the data using functional causal model (Hoyer et al., 2009). For each variable v, we samples values of v following the equation:

$$v = f(Pa_v) + \epsilon_v,$$

where f denotes the deterministic function and ϵ_v denotes the randomness. There are two kinds of equations attached to causal links between variables in $X \cup Z$. The one is linear regression with parameters drawing from uniform distribution $\mathcal{U}(-3, 3)$, and the other is non-linear neural networks with parameters drawing from uniform distribution $\mathcal{U}(-3, 3)$ and neuron numbers drawing randomly from $\{2, \cdots, 100\}$. For the non-linear function we use Relu, following (Nair and Hinton, 2010). For the causal model from X to Y, we choose equations according the values of $pa(\hat{Y}$. We modularize the values of $pa(\hat{Y}$ into several blocks: with 80% probability, we choose $f(x) = e^{-x^2}$, and with



(a) $\mathcal{G}$ **(b)** $\mathcal{G}_{\underline{X}}$

Figure 5.5: Causal Graphs.

20% probability, we choose $f(x) = \frac{1}{1+e^{-x}}$. Each value of Z is mapped into discrete space for the convenience of computation, even though our method supports continuous variables too. By doing this, we can create strongly non-linear mechanisms for generating the datasets. At the same time, we can inject ground-truth subgroups that have an exceptionally different generating mechanism compared with the other parts of the data.

Real-world Dataset For the real-world dataset, we use the Adult dataset from the UCI Machine Learning Repository (Dua and Graff, 2017). The dataset consists of 65,123 records with 14 attributes such as education, age etcetera. We choose 7 attributes and map values of Z into discrete space. For the Adult dataset, we cannot know the ground truth of data generation process. The causal graph is generated by applying the PC algorithm in Tetrad (Glymour and Scheines, 1986). For the generated causal graph, we choose Marital-status and Age as hidden confounders U_1 and U_2. Other attributes are represented as Education: Z_2, Sex: Z_1, Occupation: X_2, Hours: X_1. The associated causal graph is shown in Figure 5.5.

Baseline For experiments on both synthetic and real-world datasets, we compare our method with the method that is oblivious to causal structures. For the subgroup search process, we use (Duivesteijn et al., 2016, Algorithm 1). For the model class, we train a SVM model for each subgroup to learn the mapping function $P_{M_D}(\hat{Y}|X)$. For quality, we directly compare the performances of trained SVM models in subgroup and the whole data by computing the mean

differences of predicted samples. We call this baseline *NEMM*, distinguishing with our method *SEMM*.

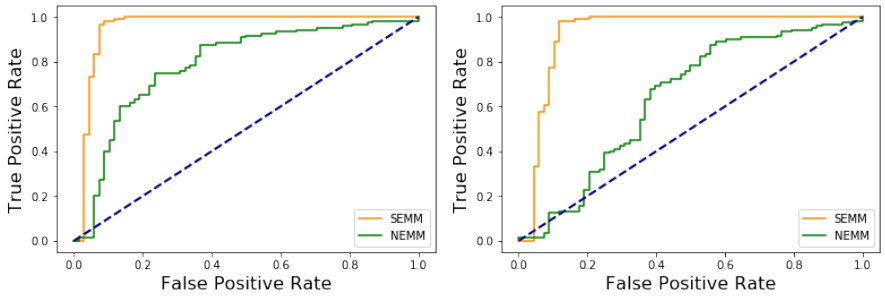
5.6.1 Experiments on Synthetic Dataset

For the experiments on synthetic data, we randomly sample 100 graphs and generate 100 datasets with equations following the above instructions. We are curious to learn whether our method finds injected exceptional subgroups, and how it performs comparing with the baseline. We run the algorithm to discover the most exceptional subgroups. Averaging results of ROC curve and AUC with standard errors, are reported in Figure 5.6a and Figure 5.7. We notice that our algorithm can reliably find the exceptional subgroups when $Q = 5$ and $Q = 10$, outperforming the baseline method that is oblivious to the causal structures. Baseline methods may discover subgroups with redundant descriptions, for which the attributes are independent of target quantities. We also notice that with the increasing Q, AUC decreased. The reason might be that with the increase of Q, subgroups with lower quality score came in, which may contain false discoveries.

In order to evaluate the efficiency of our methods against the number of attributes K and the number of records N, we fix other parameters and vary K and N respectively. By doing this, we generate 10 graphs and datasets for each K and N and report the average runtime with standard errors in Figure 5.8. We can see that the runtime for algorithm that ignores the functional constraints increases nearly exponentially with the increase of K and N. The reasons are two-fold: on the one hand, the spurious associations may cost the algorithm more time doing redundant search; on the other hand, with the increase of N, training time for the model in each subgroup would increase. Conversely, our algorithm does not need to train the model from scratch in subgroups, which leads to increased efficiency.

5.6.2 Experiments on Real-world Dataset

For the real-world dataset, we do not know the ground-truth subgroups. We propose to measure the difference between $P(\hat{Y}|X)$ in the space $\mathbb{R}$ instead of mapping values into label space. We propose to report the discoveries with p-values by the following method. For each quality score $\varphi(D)$, we can compute its p-value with the following steps: we assume that $P_{\hat{Y}|X}$ and $P^{D}_{\hat{Y}|X}$ come from the same distribution $P(P_{\hat{Y}|X})$ with regard to $P(\Omega)$. Then, empirically,



(a) Q=5 **(b)** Q=10

Figure 5.6: ROC curves for various setting of Q.

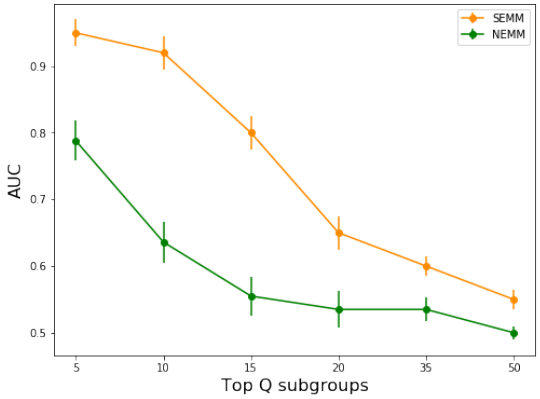
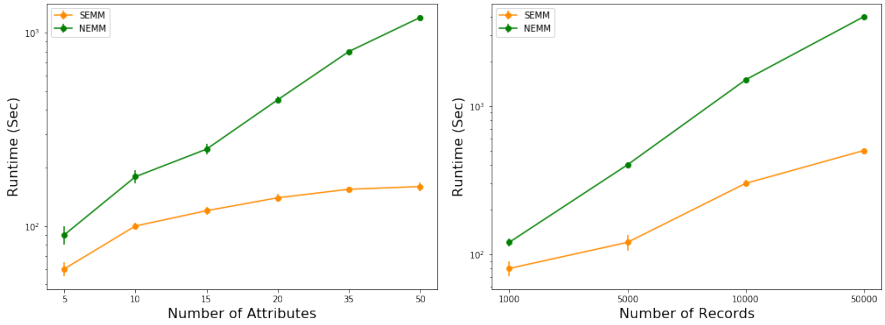


Figure 5.7: Area under ROC curve for various setting of Q.

we combine samples $\{\hat{y}\}$, $\{\hat{y}^D\}$ and randomly shuffle them by replacing elements between two sample sets. We can compute the new MMD_u^2 with the shuffled data. We repeat this ten thousand times. By doing so, we can formulate a null distribution of quality scores and get the p-value for $\varphi(D)$. If $\varphi(D)$ is so large as to be outside the 1-β quantile of the null distribution, we can reject the null hypothesis. This means we are confident that the quantity $P_{M_D}(\hat{Y}|X)$ in subgroup D is significantly different with the quantity $P_M(\hat{Y}|X)$ in the whole dataset. The top-5 exceptional subgroups are reported in Table 5.1. The reported p-values tell us that we can be confident about the exceptionality scores.



(a) with respect to K. **(b)** with respect to N.

Figure 5.8: Runtime sensitivity with respect to various parameters.

Table 5.1: Experiments on real-world datasets. Higher $\varphi_{\hat{Y}|X}(D)$ means more exceptional.

D	$\varphi_{\hat{Y}\|X}(D)$	$\frac{\|D\|}{N}$	p-value
Assoc-acdm $\wedge$ Female	$4.19 \cdot 10^{-09}$	.02	0.0004
11th	$3.57 \cdot 10^{-09}$	.04	0.0026
Assoc-voc	$3.29 \cdot 10^{-09}$	.05	0.0034
Bachelors $\wedge$ Female	$2.57 \cdot 10^{-09}$	.05	0.0131
HS-grad $\wedge$ Male	$1.81 \cdot 10^{-09}$	.21	0.0358

5.7 Conclusion

In this chapter, we study a general problem: how to find the most exceptional subgroups w.r.t differences between statistical quantity $P(\hat{Y}|X)$ in the whole data and in the subgroups. We argue that exceptional model mining / subgroup discovery should consider the underlying data generation mechanism. We propose D-graph, a causal graph with extra nodes pointing to descriptive variables, which indicate the change of local mechanisms. We further propose to find and leverage functional constraints to boost the subgroup search process, w.r.t associated causal graph. We propose an information theoretic quality measure, to estimate the difference between $P(\hat{Y}|X)$.

Experiments on synthetic and real-world datasets are conducted to evaluate whether our method can discover exceptional subgroups reliably, effectively and efficiently. The experiment results show that our methods can significantly outperform the baseline which ignores causal mechanism. On synthetic data,

our method outperforms the causality-oblivious baseline in terms of AUC over a range of top-Q EMM tasks with varying Q (cf. Figure 5.7). Moreover, the outperformance goes beyond simple AUC measurements: as Figure 5.6 displays: the entire ROC curve for our method dominates the ROC curve for the baseline, i.e., for any choice of False Positive Rate (FPR), our method has a True Positive Rate (TPR) that is either equally good as or better than the baseline. An added benefit is provided in terms of runtime: taking causality into account during the search process, and restricting the search to descriptions that satisfy Minimality (cf. Corollary 5.5.1), ensures that the runtime is much less sensitive to parameters K, and N (cf. Figure 5.8; notice that the y-axis is in logspace): especially with respect to K, the number of attributes in the graph, the baseline shows an exponential increase in runtime while the runtime curve for our algorithm flattens off in logspace. Additional experiments on real-world data (cf. Table 5.1) illustrate that our method can find statistically significantly exceptional causal subgroups, where significance is derived from a permutation test.

6
Conclusion

"Mathematics needs both birds and frogs. Mathematics is rich and beautiful, because birds give it broad visions and frogs give it intricate details. Mathematics is both great art and important science, because it combines generality of concepts with depth of structures."

Birds and Frogs,
Freeman Dyson, 2009.

In this dissertation, we studied the problem of uncertainty in Exceptional Model Mining. EMM is a powerful data mining framework that allows us to discover cohesive subsets from the whole dataset, in which the interactive patterns between target variables are exceptional, compared with those interactive patterns in the whole dataset. Because of the interpretable descriptions associated with the subgroups, the EMM framework can provide additional values for the study of fairness and explainable methods. However, all these applications have to be built upon knowing the reliability of the discoveries. This requires the study of uncertainty in EMM.

We investigated the uncertainty in EMM by studying the underlying mechanisms that determine the exceptionality score of each subgroup. The general process of computing the exceptionality score consists of the following steps: for a given description language and a dataset at hand, we can formulate subgroups in terms of attribute variables; the cohesive records covered by a subgroup can be used to learn a model by the pre-defined model class; then a quality measure is employed to map the performance of the model to a real-valued quality score; finally a search algorithm is used to find the top-Q exceptional subgroups guided by the quality score. In this process, the sources of uncertainty might be included in the dependency modeling and subgroup selection process. The observational data are usually imperfect with imbalanced feature distribution and missing information. Hence, learning the true interactive pat-

terns with limited data at hand, especially with limited data in subgroups, is challenging. This challenge brings uncertainty to the capture of performance for a given model and further influences the evaluation of its exceptionality. We focused on analyzing the uncertainty in dependency modeling by proposing probabilistic methods to model and infer the interactive patterns between targets. In particular, we studied the following kinds of dependency modeling with practical applications:

- *Multi-modal dependency in spatio-temporal data*. The behavioral patterns in spatio-temporal social posts are represented by distributions of spatial locations, time and word topics. Specific deviations across any combination of these three distributions can indicate interesting, exceptional behavior of the population. Properly capturing the uncertainty in these multi-modal interactions can greatly benefit EMM for finding meaningful exceptional behavioral spatio-temporal patterns. Due to the complexity of multi-modal interactions, it is difficult to estimate the interactive patterns with limited data in subgroups. Hence, we proposed to explicitly model the underlying data generating process by a Bayesian non-parametric modeling method. The quality measure based on comparing posterior distributions can give us more confidence about the exceptionality of subgroups and avoid false discoveries.
- *Heterogeneous and high-dimensional dependency in educational and network data*. Unknown heterogeneity across the data can lead a model to be very effective for some subpopulations and ineffective for some other subpopulations. The heterogeneity and complex interactions could bring extra uncertainty to the dependency modeling as well as the evaluation of exceptionality. Network representation model is an effective method that can extract and summarize such heterogeneity from high-dimensional interactions. However, in order to capture the exceptionality of such heterogeneous and high-dimensional dependencies, new quality measures are required. We proposed a new quality measure called Mean Latent Similarity Discrepancy (MLSD) based on the U-statistic. With this measure, we are able to quantify the difference between performance of network representation models. We further proposed a hypothesis testing method to validate the discoveries derived with the guidance of this quality measure against false discoveries. We employed this framework to analyze the fairness in the network representation model, which can provide a new view for fairness from the aspect of unsupervised sensitive attributes.
- *Causal dependency in observational data*. We studied a new kind of

dependency for EMM. Such dependency cannot be learned from the observational data with the existing association modeling because of the confounding bias and unobserved counterfactuals. The uncertainty in causal dependency modeling could bring new challenges for the study of uncertainty in EMM. In this context, we focused on estimating the causal dependency with observational data. In particular, we proposed a neural network framework to estimate the causal effects of a binary treatment variable. We call this framework **A**dversarial **B**alancing-based representation learning for **C**ausal **E**ffect **I**nference (**ABCEI**). ABCEI used adversarial learning to balance the distributions of treatment and control group in the latent representation space, without any assumption on the form of the treatment selection/assignment function. ABCEI preserved useful information for predicting causal effects under the regularization of a mutual information estimator. The experimental results showed that ABCEI is robust against treatment selection bias, and matches/outperforms the state-of-the-art approaches.

- Based on the study of causal dependency, we started investigating the causal dependency within and without subgroups. In particular, we studied the effect of subgroup selection on the statistical quantity of interest. We call this causal effects Local Causal Dependency. The main disadvantage for current interpretable methods is that they only consider the correlation between features and the decision making process. Just because variables are strongly associated, does not mean their relation is also causal. The reason might be that they are confounded by a third part of variables. The integration of Local Causal Dependency and EMM allows us to understand the determining mechanisms behind the performance of a model. This can prevent us from being misled by spurious associations between features and the decision making process, which can leverage our research on explainable machine learning from the level of correlations to the level of causations.

As a data driven framework, EMM is able to uncover the interesting regions in the data space where the data generating process might be exceptionally different with other regions. Studying the uncertainty in EMM can help us understand the underlying mechanism of decision models. In future work, we are going to investigate how to make use of the meta information discovered by EMM to improve machine learning models. Based on the contributions of our current work, we are particularly interested in the following work:

- For learning behavior analysis, we would like to make use of these discovered exceptional behavioral patterns to establish an ensemble model,

which can model both normal and exceptional learning behaviors for the students in MOOCs. We plan to develop a prediction model that can perform well on each part of the dataset, including the exceptional ones.

- For fairness in network representation model, we would like to integrate the representation learning and subgroup discovery into a unified framework. By doing this, we are aiming to generate fair and informative node representations for downstream applications like fair allocation or fair demands analysis.
- For causal effect inference, we would like to explore more connections between relevant methods in domain adaptation (Daume III and Marcu, 2006) and counterfactual learning (Swaminathan and Joachims, 2015b) with the methods in causal inference. A proper extension would be to consider multiple treatment assignments or the existence of hidden confounders.
- For Local Causal Dependency modeling, we would like to make use of the functional constraints to derive a causal mechanism disentangling method. This would help us to build a generative model that considers the true generating factor instead of spurious associations.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, 12(1):307–328, 1996.

Mohammad Al Hasan and Mohammed J Zaki. Output space sampling for graph patterns. *Proceedings of the VLDB Endowment*, 2(1):730–741, 2009.

Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1), 1994.

Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *arXiv preprint arXiv:1711.04710*, 2017.

Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

Martin Atzmueller and Frank Puppe. Sd-map–a fast algorithm for exhaustive subgroup discovery. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 6–17. Springer, 2006.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

Vladimir Batagelj and Ulrik Brandes. Efficient generation of large random networks. *Physical Review E*, 71(3), 2005.

Aimene Belfodil, Adnene Belfodil, and Mehdi Kaytoue. Anytime subgroup discovery in numerical domains with guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 500–516. Springer, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540. PMLR, 10–15 Jul 2018.

Ahmed Anes Bendimerad, Marc Plantevit, and Céline Robardet. Unsupervised exceptional attributed sub-graph mining in urban data. In *2016 IEEE International Conference on Data Mining*, 2016.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

Michael R Berthold, Katharina Morik, Arno Siebes, Bruno Crémilleux, Arnaud Soulet, Bernd Wiswedel, Michael Wurst, Niall Adams, David J Hand, Elisa Fromont, et al. Parallel universes and local patterns.

Dimitris Bertsimas, Nikita Korolko, and A Weinstein. Identifying exceptional responders in randomized trials: An optimization approach. *INFORMS Journal on Optimization, under review*, 2018.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, 2010.

Mario Boley, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner. Direct local pattern sampling by efficient two-step random procedures. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 582–590, 2011.

Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social Networks*, 21(4), 2000.

Guillaume Bosc, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue. Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data Mining and Knowledge Discovery*, 32(3):604–650, 2018.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. Relaxed functional dependencies—a survey of approaches. *IEEE Transactions on Knowledge and Data Engineering*, 28 (1):147–165, 2015.

Sabrina Casucci, Li Lin, Sharon Hewner, and Alexander Nikolaev. Estimating the causal effects of chronic disease combinations on 30-day hospital readmissions based on observational medicaid data. *Journal of the American Medical Informatics Association*, 25(6):670–678, 2017.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009.

Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 757–766. JMLR. org, 2017.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018a.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*, 2018b.

Wei Chen, Zhaosong Huang, Feiran Wu, Minfeng Zhu, Huihua Guan, and Ross Maciejewski. Vaud: A visual analysis approach for exploring spatio-temporal urban data. *IEEE Transactions on Visualization and Computer Graphics*, 24(9):2636–2648, 2018c.

Flavio Chierichetti, Jon M Kleinberg, Ravi Kumar, Mohammad Mahdian, and Sandeep Pandey. Event detection via communication pattern analysis. In *Proc. ICWSM*, pages 51–60, 2014.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

YooJung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy Van den Broeck. Learning fair naive Bayes classifiers by discovering and eliminating discrimination patterns. *arXiv preprint arXiv:1906.03843*, 2019.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Juan D Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 119–128. ACM, 2010.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. In *Advances in Neural Information Processing Systems*, pages 4655–4665, 2018.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, 1983.

Vincent Dorie. Npci: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci*, 2016.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Xin Du, Wouter Duivesteijn, Martijn Klabbers, and Mykola Pechenizkiy. ELBA: Exceptional learning behavior analysis. *International Educational Data Mining Society*, 2018.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *arXiv preprint arXiv:1904.13335*, 2019.

Xin Du, Wouter Duivesteijn, Jin Tian, and Mykola Pechenizkiy. Why does my model perform differently? When exceptional model mining meets causal graph. *under review*, 2020a.

Xin Du, Yulong Pei, Wouter Duivesteijn, and Mykola Pechenizkiy. Exceptional spatio-temporal behavior mining through Bayesian non-parametric modeling. *Data Mining and Knowledge Discovery*, pages 1–24, 2020b.

Xin Du, Yulong Pei, Wouter Duivesteijn, and Mykola Pechenizkiy. Fairness in network representation by latent structural heterogeneity in observational data. In *34th AAAI conference on Artificial Intelligence (AAAI2020)*, 2020c.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Sumeet Dua, U Rajendra Acharya, and Prerna Dua. *Machine learning in healthcare informatics*, volume 56. Springer, 2014.

Wouter Duivesteijn and Arno Knobbe. Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. In *2011 IEEE International Conference on Data Mining*, 2011.

Wouter Duivesteijn and Julia Thaele. Understanding where your classifier does (not) work–the scape model class for emm. In *2014 IEEE International Conference on Data Mining*, pages 809–814. IEEE, 2014.

Wouter Duivesteijn, Arno Knobbe, Ad Feelders, and Matthijs van Leeuwen. Subgroup discovery meets bayesian networks–an exceptional model mining approach. In *2010 IEEE International Conference on Data Mining*, pages 158–167. IEEE, 2010.

Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. ITCS*, pages 214–226. ACM, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Arnaud Giacometti and Arnaud Soulet. Dense neighborhood pattern sampling in numerical data. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 756–764. SIAM, 2018.

Fosca Giannotti, Lorenzo Gabrielli, Dino Pedreschi, and Salvatore Rinzivillo. Understanding human mobility with big data. In *Solving Large Scale Learning Tasks. Challenges and Algorithms*, pages 208–220. Springer, 2016.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 2002.

Clark Glymour and Richard Scheines. Causal modeling with the tetrad program. *Synthese*, 68 (1):37–63, 1986.

Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 1*, ICCV '03, pages 487–493, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-1950-4.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar), 2012.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.

Peter D Grünwald and Abhijit Grunwald. *The minimum description length principle*. MIT press, 2007.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

David J Hand. Pattern detection and discovery. In *Pattern Detection and Discovery*, pages 1–12. Springer, 2002.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Daniel E Ho, Kosuke Imai, Gary King, Elizabeth A Stuart, et al. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, pages 769–778. ACM, 2012.

Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 495–503. SIAM, 2016.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Martin Jankowiak and Manuel Gomez-Rodriguez. Uncovering the spatiotemporal patterns of collective social activity. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 822–830. SIAM, 2017.

Dominik Janzing, Kailash Budhathoki, Lenon Minorics, and Patrick Blöbaum. Causal structure based root cause analysis of outliers. *arXiv preprint arXiv:1912.02724*, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 652–661. JMLR. org, 2016.

Ruoming Jin, Victor E Lee, and Hui Hong. Axiomatic ranking of network role similarity. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8908–8919, 2018a.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*, 2018b.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Mohammad Khajah, Rowan Wing, Robert Lindsey, and Michael Mozer. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Educational Data Mining 2014*, 2014.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Kyoung-Sook Kim, Isao Kojima, and Hirotaka Ogawa. Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science*, 30(9):1899–1922, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, page 201309933, 2014.

Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. 1996.

Konstantin Knauf, Daniel Memmert, and Ulf Brefeld. Spatio-temporal convolution kernels. *Machine Learning*, 102(2):247–273, 2016.

Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.

Sotiris Kotsiantis, E Koumanakos, D Tzelepis, and V Tampakas. Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence*, 3 (2):104–110, 2006.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1617–1626. ACM, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

Andrea Lamont, Michael D Lyons, Thomas Jaki, Elizabeth Stuart, Daniel J Feaster, Kukatharmini Tharmaratnam, Daniel Oberski, Hemant Ishwaran, Dawn K Wilson, and M Lee Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in Medical Research*, 27(1):142–157, 2018.

Nicholas D Lane, Li Pengyu, Lin Zhou, and Feng Zhao. Connecting personal-scale sensing and networked community behavior to infer human activities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 595–606. ACM, 2014.

Nada Lavrač, Peter Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *International Conference on Inductive Logic Programming*, pages 174–185. Springer, 1999.

Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 1–16. Springer, 2008.

Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Mining subgroups with exceptional transition behavior. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 965–974, 2016.

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168. IEEE, 2011.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pages 929–939, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proc. FAT**, pages 349–358. ACM, 2019.

John F Mahoney and James M Mohen. Method and system for loan origination and underwriting, October 23 2007. US Patent 7,287,008.

Michael Mampaey, Siegfried Nijssen, Ad Feelders, and Arno Knobbe. Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data. In *2012 IEEE 12th International Conference on Data Mining*, pages 499–508. IEEE, 2012.

Michael Mampaey, Siegfried Nijssen, Ad Feelders, Rob Konijn, and Arno Knobbe. Efficient algorithms for finding optimal binary features in numeric and nominal labeled data. *Knowledge and Information Systems*, 42(2):465–492, 2015.

Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering reliable approximate functional dependencies. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 355–363, 2017.

Marvin Meeng, Wouter Duivesteijn, and Arno Knobbe. Rocsearch—an roc-guided search strategy for subgroup discovery. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 704–712. SIAM, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013b.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Stephen L Morgan and David J Harding. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, 35(1):3–60, 2006.

Katharina Morik, Jean-François Boulicaut, and Arno Siebes. Local pattern detection: International seminar dagstuhl castle, germany, april 12-16, 2004, revised selected papers (lecture notes in computer science/lecture notes in artificial intelligence), 2005.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 1997.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, University of British Columbia, 2007.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

Solomon Negash and Paul Gray. Business intelligence. In *Handbook on decision support systems 2*, pages 175–193. Springer, 2008.

Yang Ning, Sida Peng, and Kosuke Imai. Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv preprint arXiv:1812.08683*, 2018.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

Michal Ozery-Flato, Pierre Thodoroff, and Tal El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence*, pages 97–110. Springer, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3), 2017.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. 2017.

Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Spatio-temporal random fields: compressible representation and distributed estimation. *Machine Learning*, 93(1):115–139, 2013.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008.

Kai Puolamäki, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. Interactive visual data exploration with subjective feedback. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 214–229. Springer, 2016.

Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. Modeling and predicting learning behavior in moocs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 93–102. ACM, 2016.

Minghui Qiu, Feida Zhu, and Jing Jiang. It is not just what we say, but how we say them: LDA-based behavior-topic model. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 794–802. SIAM, 2013.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188, 2001.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Daniel T Seaton, Yoav Bergner, Isaac Chuang, Piotr Mitros, and David E Pritchard. Who does what in a massive open online course? *Communications of the ACM*, 57(4):58–65, 2014.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Kijung Shin, Tina Eliassi-Rad, and Christos Faloutsos. Patterns and anomalies in k-cores of real-world graphs with applications. *Knowledge and Information Systems*, pages 1–34, 2017.

Dominique T Shipmon, Jason M Gurevitch, Paolo M Piselli, and Stephen T Edwards. Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data. *arXiv preprint arXiv:1708.03665*, 2017.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Jeffrey A Smith and Petra E Todd. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353, 2005.

Lindsay I Smith. A tutorial on principal components analysis. Technical report, 2002.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

Joram Soch and Carsten Allefeld. Kullback-leibler divergence for the normal-gamma distribution. *arXiv preprint arXiv:1611.01437*, 2016.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in Medicine*, 35(16):2741–2753, 2016.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Masashi Sugiyama and Matthias Krauledat. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Lei Sun and Alexander G Nikolaev. Mutual information based matching for causal inference with observational data. *The Journal of Machine Learning Research*, 17(1):6990–7020, 2016.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1): 1731–1755, 2015a.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015b.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *AAAI/IAAI*, pages 567–573, 2002a.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 519–527. Morgan Kaufmann Publishers Inc., 2002b.

Hannu Toivonen. Sampling large databases for association rules. volume 96, pages 134–145, 1996.

Stephen Tu. The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. Technical report, Computer Science Division, UC Berkeley, 2014.

Matthijs van Leeuwen and Arno Knobbe. Non-redundant subgroup discovery in large and complex data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer, 2011.

Matthijs van Leeuwen and Arno Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2):208–242, 2012.

Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017.

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

Feng Wang and Li Chen. A nonlinear state space model for identifying at-risk students in open online courses. *EDM*, 16:527–532, 2016.

Lisa Wang, Angela Sy, Larry Liu, and Chris Piech. Learning to represent student knowledge on programming exercises using deep learning. In *Proceedings of the 10th International Conference on Educational Data Mining; Wuhan, China*, pages 324–329, 2017.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.

Xian Wu, Yuxiao Dong, Chao Huang, Jian Xu, Dong Wang, and Nitesh V Chawla. Uapd: Predicting urban anomalies from spatial-temporal data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 622–638. Springer, 2017.

Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831. ACM, 2012.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2634–2644, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Quan Yuan, Wei Zhang, Chao Zhang, Xinhe Geng, Gao Cong, and Jiawei Han. Pred: Periodic region detection for mobility modeling of social media users. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 263–272. ACM, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representation. *arXiv preprint arXiv:1906.08386*, 2019.

Siyuan Zhao and Neil Heffernan. Estimating individual treatment effects from educational studies with residual counterfactual networks. In *10th International Conference on Educational Data Mining*, 2017.

Zhe Zhao, Zhiyuan Cheng, Lichan Hong, and Ed H Chi. Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1406–1416. International World Wide Web Conferences Steering Committee, 2015.

Xin Zheng, Jialong Han, and Aixin Sun. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018.

Yixian Zheng, Wenchao Wu, Yuanzhe Chen, Huamin Qu, and Lionel M Ni. Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data*, 2(3):276–296, 2016.

Yu Zheng, Huichu Zhang, and Yong Yu. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 2. ACM, 2015.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

List of Figures

List of Tables

List of Acronyms

ABCEI	Adversarial Balancing-based representation learning for Causal Effect Inference
ATE	Average Treatment Effect
ATT	Average Treatment Effect on Treated
AUC	Area Under ROC Curve
BART	Bayesian Additive Regression Trees
BLR	Balancing Linear Regression
BNN	Balancing Neural Network
BNPM	Bayesian Non-Parametric Model
CATE	Conditional Average Treatment Effect
C-Component	Confounded Component
CEVAE	Causal Effect Variational Autoencoders
CF	Causal Forests
CFR-Wass	Counterfactual Regression with Wasserstein Distance
CRP	Chinese Restaurant Process
DNN	Deep Neural Networks
DV	Donsker-Varadhan
ELU	Exponential Linear Unit
EMM	Exceptional Model Mining
FPR	False Positive Rate
FD	Functional Dependency
FIM	Frequent Itemset Mining
GAN	Generative Adversarial Network
IPM	Integral Probability Metric
IPS	Inverse Propensity Score
ITE	Individual Treatment Effect
KNN	K-Nearest Neighbor
LCD	Local Causal Dependency
LSTM	Long Short-Term Memory
LPM	Local Pattern Mining
MDL	Minimum Description Length
MOOCs	Massive Open Online Courses
MI	Mutual Information
MLE	Maximum Likelihood Estimation
MLSD	Mean Latent Similarity Discrepancy
MMD	Maximum Mean Discrepancy

PCA	Principle Component Analysis
PEHE	Precision in Estimation of Heterogeneous Effect
RCT	Randomized Controlled Trials
RF	Random Forests
RKHS	Reproducing Kernel Hilbert Space
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SD	Subgroup Discovery
SCM	Structural Causal Model
SVM	Support Vector Machine
PO	Potential Outcome Framework
TARNet	Treatment-Agnostic Representation Networks
TPR	True Positive Rate

Acknowledgments

I feel fortune to have the chance to pursue my PhD degree at Technische Universiteit Eindhoven in the Netherlands. The pursuing process makes me grown up and changes my lifestyle. This is partly because of the PhD study itself, and more importantly, because of the people that I met during the journey. I would like to express my appreciate to them in this place with several paragraphs.

Firstly, I would like to thank my supervisors, prof. dr. Mykola Pechenizkiy and dr. Wouter Duivesteijn. Thank you for your guidance and support. There is an ancient quote: 'Talents are everywhere, however seldom can they be recognized.' I still remember the day when Mykola came to Delft with George and we had a sushi dinner together. That dinner gave me a chance to pursue my PhD degree in Computer Science. During my PhD study, Mykola and Wouter had shown great support to my research. I can have a lot of freedom to explore what I am interested in. When I met problems during my study, they always encourage me to follow my way, which helped me a lot to improve myself with full confidence. I remember in DBDBD 2017 after I had a bad presentation, Wouter posted on social media saying that I did a good job. This encouragement helped me to gain confidence and achieve better performance in the next time. These are only small parts of the stories about how my supervisors helped me in my study. I would like to thank Wouter for the group lunch in every working day and for the social events such as Poker night. Also many thanks to Mykola for holding the group activities like annual cycling trip, bowling events and cooking events. These events really helped me a lot to adapt to the lifestyle in the Netherlands. I would also like to thank Mykola for offering me another job as a scientific programmer in the department. Without this position, I cannot even have the chance to finish my PhD.

I would like to thank the committee members for this PhD defense, for carefully reviewing this dissertation and for the insightful feedbacks: prof. dr. Dino Pedreschi from Università di Pisa, prof. dr. Alexandre Termier from Université de Rennes 1, dr. Alessandro Di Bucchianico from Technische Universiteit Eindhoven and dr. Marc Plantevit from Université Claude Bernard Lyon 1. I would also like to thank my adviseur dr. Esther Galbrun from University of Eastern Finland. Your insightful and detailed feedbacks helped me to make this dissertation more complete.

I would like to thank my co-authors dr. Lei Sun, dr. Alexander Nikolaev,

prof. dr. George Fletcher, dr. Sibylle Hess, dr. Yulong Pei, dr. Jianpeng Zhang, dr. Vlado Menkovski, dr. Hao Wang and Yuhao Wang. I have learned a lot by working with you during our collaboration and I feel fortune to have our names together on the scientific publications.

In June 2019 I had an opportunity to shortly visit the exploratory data analysis group at University of Helsinki in Finland. I would like to thank dr. Kai Puolamäki and dr. Emilia Oikarinen for their warm welcome. During the short academic visiting we had several discussions on a couple of interesting research problems. I have learned a lot from those discussions and from people in their research team. I would also like to thank them for the delicious dinner.

I would like to thank Simon van der Zon, Iftitahu Nimah and Oren Zeev Ben Mordehai for their warm welcome in the first office that we shared together. There were a lot of good memories when we were working together in that office. I would also like to thank dr. Yulong Pei, Shiwei Liu, Yuhao Wang, Xuming Meng, Loek Tonnaer and Marijn Knippenberg for sharing the same office. There were a lot of discussions in our office, not only academic topics, but also joys and burdens of the PhD life. Special thanks to Loek Tonnaer, Marijn Knippenberg, Shiwei Liu and Simon van der Zon, as a team we won the first place together during the pub quiz event in our department.

I would like to thank dr. Jianpeng Zhang, dr. Cong Liu, dr. Yulong Pei, dr. Guangming Li and dr. Long Cheng for their warm introducing when I came to this group. I would also like to thank them and other friends including Shiwei Liu, Tianjin Huang, Yuhao Wang, Kaijie Zhu, Fang Lyu, Xuming Meng, Lu Yin, Hao Li, Jun Lin, Zhaohuan Wang and Mengting Jiang for organizing and joining the social events like movie night, BBQ and Hotpot parties.

I would like to express my special thanks for José, Ine and Riet. Thanks for your kind services which make our working environment more convenient. Also special thanks to José for doing proof reading for my thesis and holding Dutch games at home. We shared the joyful and warming parties. During the party, I learned 'Life is a party, but you must self the slingers uphanging!'.

My gratitude goes to prof. Paul De Bra for leading the Web Engineer group. The puzzles you made brought a lot of fun for our coffee breaks. I would also like to thank dr. Anil Yaman, Pieter Gijsbers, Bilge Celik, Simon van der Zon and Vlado Menkovski for the bouldering events. I learned how to make a master plan and focus on the movement of the body during the climbing. I would like to thank Ricky Fajri, dr. Jianpeng Zhang, Simon van der Zon and prof. dr. Mykola Pechenizkiy for sharing the house during our trip to KDD conference in London at 2018. That was a nice trip sharing both a

lot of joys and knowledge. I miss the moment when we were cooking pancake together after the whole day's conference. I would also like to thank Wouter for taking me to the ice hockey match during our trip to AAAI in New York. That was a wonderful memory. I would also like to thank Simon van der Zon, Anil Yaman, Thomas Mulder, Pieter Gijsbers, Prabhant Singh, Hamid Shahrivari Joghan, Loek Tonnaer and Marijn Knippenberg for the video game nights. We shared exciting time by playing RTS video games together.

I would also like to thank other colleagues from Data Mining group including Decebal Mocanu, Joaquin Vanschoren, Sibylle Hess, Samaneh Khoshrou, Marcos de Paula Bueno, Munehiro Kitaura, Sahithya Ravi, Prabhant Singh, Hilde Weerts, Simon Koop, Rianne Schouten, Ghada Sokar; colleagues from Database System group including George Fletcher, Odysseas Papapetrou, Nikolay Yakovets, Wilco van Leeuwen, Daphne Miedema; and colleagues from UAI group including Cassio de Campos, Robert Peharz, Alvaro Correia. Thanks for joyful and inspiring talks during lunch and coffee breaks.

I would like to thank dr. Qiang Liu, Rendong Liu and Simin Chen for maintaining the 'Dog group' together, a group initiated by the purpose of constructing a startup project but end up with holding hotpot parties. I would also like to thank dr. Qiang Liu, dr. Xiang Fu, dr. Yue Zhao and dr. Jian Fang for maintaining the 'Big brothers (Da lao men) except me group' together, a group initiated by the purpose of playing snowboard. This time we managed to do some snowboard together. I would like to express my special thanks to my previous colleagues at 3D geoinformation group in Delft including prof. dr. Jantien Stoter, dr. Hugo Ledoux, dr. Ravi Peters, dr. Ken Arroyo Ohori, Kavisha, dr. Filip Biljecki, dr. Fangyu Li, dr. Shuangfeng Wei, dr. Liu Liu and dr. Zhiyong Wang for their warm welcome and introducing when I just came to the this country. I would also like to thank dr. Ding Ding, dr. Mengshi Yang, dr. Zhijun Wang, Xiangzhen Kong, Mei Liu, Jiani Liu, dr. Riming Wang, dr. Hao Yu, Tiantian Du, Meng Meng, Taozhi Zhuang, Ran Xiao, Juan Yan, Yan Liu, Sihang Qiu, Zhenwu Wang, Yan Song, Yuxi Liu, Hongjuan Wu, Yan Teng, Rui Li, Xiaogeng Ren, Tianchen Dai, Xing Zheng, Li Lyu, Pengling Zhu and Kaiyi Zhu for sharing joyful moments in Delft.

Last but not least, I would like to express my deepest gratitude to my dear parents. You are always so supportive and understanding throughout my life. You always show your love and kindness to help me in my study.

Xin Du
Eindhoven, August 2020

Curriculum Vitæ

Xin Du was born on 24-10-1987 in Hebi, China. He received bachelor's degree in Geographic Information System from Yunnan University in 2010. Then he received his master's degree in Cartography and Geographic Information System from Wuhan University in 2015. From 2017 he started a PhD project in the Department of Mathematics and Computer Science at Eindhoven University of Technology under the supervision of dr. Wouter Duivesteijn and prof.dr. Mykola Pechenizkiy at Eindhoven, the Netherlands of which the results are presented in this dissertation.

Publications

Publications related to this dissertation:

5. **Du, Xin and Duivesteijn, Wouter and Tian, Jin and Pechenizkiy, Mykola**, *Why Does My Model Perform Differently? When Exceptional Model Mining Meets Causal Graph*, Under review, 2020.

4. **Du, Xin and Pei, Yulong and Duivesteijn, Wouter and Pechenizkiy, Mykola**, *Exceptional spatio-temporal behavior mining through Bayesian non-parametric modeling*, Data Mining and Knowledge Discovery **1–24**, (2020), Springer.

3. **Du, Xin and Pei, Yulong and Duivesteijn, Wouter and Pechenizkiy, Mykola**, *Fairness in network representation by latent structural heterogeneity in observational data*, 34th AAAI conference on Artificial Intelligence, (2020), AAAI.

2. **Du, Xin and Sun, Lei and Duivesteijn, Wouter and Nikolaev, Alexander and Pechenizkiy, Mykola**, *Adversarial balancing-based representation learning for causal effect inference with observational data*, Under review at Data Mining and Knowledge Discovery, arxiv, (2019), Preprint.

1. **Du, Xin and Duivesteijn, Wouter and Klabbers, Martijn and Pechenizkiy, Mykola**, *ELBA: Exceptional Learning Behavior Analysis*, Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018).

Publications unrelated to this dissertation:

4. **Duivesteijn, Wouter and Hess, Sibylle and Du, Xin**, *How to Cheat the Page Limit*, WIREs Data Mining and Knowledge Discovery, (2020).

3. **Pei, Yulong and Du, Xin and Zhang, Jianpeng and Fletcher, George and Pechenizkiy, Mykola**, *struc2gauss: Structure Preserving Network Embedding via Gaussian Embedding*, Data Mining and Knowledge Discovery, (2020).

2. **Wang, Yuhao and Menkovski, Vlado and Wang, Hao and Du, Xin and Pechenizkiy, Mykola**, *Causal Discovery from Incomplete Data: A Deep Learning Approach*, Under review, ArXiv preprint, (2020).

1. **Pei, Yulong and Du, Xin and Zhang, Jianpeng and Fletcher, George and Pechenizkiy, Mykola**, *Dynamic Network Representation Learning via Gaussian Embedding*, Graph Representation Learning Workshop (NeurIPS), (2019).

SIKS Dissertations

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models
02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems
04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference
05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage
07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction
08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues
09 Tim de Jong (OU), Contextualised Mobile Media for Learning
10 Bart Bogaert (UvT), Cloud Content Contention
11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective
12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining
13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling
14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets
15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval
16 Maarten Schadd (UM), Selective Search in Games of Different Complexity
17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness
18 Mark Ponsen (UM), Strategic Decision-Making in complex games
19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles
20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems
22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection
46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns

06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning
43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data
02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method
03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions
04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation
05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability
06 Damian Tamburri (VU), Supporting Networked Software Development
07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints
09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language
10 Ivan Salvador Razo Zapata (VU), Service Value Networks
11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support
12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control
13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains
14 Yangyang Shi (TUD), Language Models With Meta-information
15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software

30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

---

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores

	28	Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
	29	Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
	30	Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
	31	Yakup Koç (TUD), On the robustness of Power Grids
	32	Jerome Gard (UL), Corporate Venture Management in SMEs
	33	Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
	34	Victor de Graaf (UT), Gesocial Recommender Systems
	35	Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction
2016	01	Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
	02	Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
	03	Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
	04	Laurens Rietveld (VU), Publishing and Consuming Linked Data
	05	Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
	06	Michel Wilson (TUD), Robust scheduling in an uncertain environment
	07	Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
	08	Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
	09	Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
	10	George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
	11	Anne Schuth (UVA), Search Engines that Learn from Their Users
	12	Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
	13	Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
	14	Ravi Khadka (UU), Revisiting Legacy Software System Modernization
	15	Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UVA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TUE), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UVA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VU), Logics for causal inference under uncertainty
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VU), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TUE), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TUE), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs
38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TUE), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 **Xin Du (TUE), The Uncertainty in Exceptional Model Mining**