Explaining Deviating Subsets through Explanation Networks

Antti Ukkonen[1], Vladimir Dzyuba[2], and Matthijs van Leeuwen[3]

[1] Department of Computer Science, University of Helsinki, Helsinki, Finland,
antti.ukkonen@gmail.com
[2] Department of Computer Science, KU Leuven, Leuven, Belgium,
vladimir.dzyuba@cs.kuleuven.be
[3] LIACS, Leiden University, Leiden, The Netherlands,
m.van.leeuwen@liacs.leidenuniv.nl

Abstract. We propose a novel approach to finding explanations of deviating subsets, often called *subgroups*. Existing approaches for subgroup discovery rely on various quality measures that nonetheless often fail to find subgroup sets that are diverse, of high quality, and most importantly, provide good explanations of the deviations that occur in the data.
To tackle this issue we introduce *explanation networks*, which provide a holistic view on all candidate subgroups and how they relate to each other, offering elegant ways to select high-quality yet diverse subgroup sets. Explanation networks are constructed by representing subgroups by nodes and having weighted edges represent the extent to which one subgroup explains another. Explanatory strength is defined by extending ideas from database causality, in which interventions are used to quantify the effect of one query on another.
Given an explanatory network, existing network analysis techniques can be used for subgroup discovery. In particular, we study the use of PageRank for pattern ranking and seed selection (from influence maximization) for pattern set selection. Experiments on synthetic and real data show that the proposed approach finds subgroup sets that are more likely to capture the generative processes of the data than other methods.

1 Introduction

Within the field of exploratory data mining, subgroup discovery (SD) [7, 21] is concerned with finding and explaining deviating subsets, i.e., regions in the data that stand out with respect to a given target. It has a number of closely related cousins, such as significant pattern mining [20] and emerging pattern mining [3], which all concern the discovery of patterns correlated with a Boolean target concept. The subgroup discovery task is more generic, as it is agnostic of the data and pattern types. For example, the target could be discrete or numeric [13], both of which we consider in this paper.

Since its introduction twenty years ago, many algorithms and quality measures have been proposed in the literature. While the initial focus was on devising

more efficient algorithms, over time the focus has shifted towards redundancy elimination [9, 14, 10], (statistical) validation [5], and generalization of the task [12]. Nevertheless, existing approaches have several limitations, in particular where it concerns the core of the subgroup discovery task: *providing accurate explanations of the deviations that occur in the data*. This has several causes.

First of all, quality measures in subgroup discovery traditionally combine—often by multiplication—the size of the subgroup, i.e., the number of rows it covers, with its effect, i.e., the extent to which the target value for those rows deviates from the dataset average. The problem with this approach is that this results in a somewhat arbitrary trade-off between size and effect that has a very large impact on the scores (and thus ranking) of all patterns.

Second, most approaches that aim to eliminate redundancy take these individual qualities for granted and primarily consider the covers of subgroups to discard redundant patterns. This is true for, e.g., approaches based on relevancy/closedness [9] and generalization-aware pruning [14]. Furthermore, none of these methods explicitly considers redundancy among subgroups that do not share any attributes among their descriptions, neither do they explicitly consider the possibility that one description may be more interesting/relevant than another. That is, a subgroup is either kept or discarded and no alternatives are offered, where the exact choice between similar subgroups is pretty much random. Some approaches, such as DSSD [10], are heuristic and defined procedurally, which makes it even harder to assess the results.

As a consequence of the above, existing methods do not provide accurate *descriptions* for deviating subsets in practice, as we will empirically show in Section 6. That is, finding deviating subsets and descriptions that correspond to those subsets is doable, but identifying *accurate explanations*, i.e., descriptions that capture the data generating process, is a much harder task.

Approach and contributions We introduce *explanation networks*, i.e., networks in which the nodes represent subgroups and weighted, directed edges represent *explanations* between pairs of subgroups. Explanation networks offer a global perspective on all subgroups and their relationships, regardless of the branches of the search tree the subgroups happen to reside. As a result, the network naturally represents all information concerning relevancy and redundancy.

Technically, we build on ideas from *database causality* [15, 16] to quantify to what extent subgroups "explain" each other. In particular, we use the notion of an *intervention* [22, 19]: we say that a subgroup T explains a subgroup S if removing the cover of T from the cover of S results in a (much) smaller effect. Because a larger T is more likely to (partially) explain S by chance than a smaller T, we normalize its explanatory influence with its *expected* explanation. The result is an elegant formula that quantifies explanatory strength with a number of desirable properties. For example, it explicitly distinguishes effect from cover size and accounts for both in a principled way; others will be discussed later.

We demonstrate the strengths of explanation networks through two different pattern mining tasks. First, we show how subgroups can be ranked based on their global explanatory power, i.e., by considering all pairwise relationships.

We achieve this by observing that this setting is analogue to that of identifying relevant webpages on the World Wide Web and thus apply PageRank to explanation networks. Second, we observe another analogy to network analysis and show how the pattern set selection task, i.e., the task of selecting a small and diverse set of subgroups, can be formalized as a seed selection (Influence Maximization) problem on explanation networks.

The remainder of the paper is organized as follows. First, we discuss related work in Section 2, followed by preliminaries in Section 3. We then formally introduce explanation networks in Section 4 and describe the two tasks in Section 5. Section 6 presents experiment results, both on synthetic and real data, after which we conclude with conclusions and a brief outlook in Section 7.

2 Related work

This section provides a high-level overview of two categories of related work aimed at discovering and explaining phenomena observed in data, namely 1) causality in databases, and 2) subgroup discovery.

Database causality Establishing "actual causality" requires controlled randomized experiments and thus cannot be accomplished using purely observational data [17]. Database research therefore uses a relaxed definition of causality, which originates from *database provenance* and focuses on identifying causal relations between tuples, i.e., which tuples in a database (input tuples) affect the results of a query (output tuples or columns thereof) [15, 16]?

An important extension of this line of work replaces fine-grained "causes" described by (potentially large) collections of individual tuples by coarse-grained *explanations*, i.e., concise descriptions of those collections in a certain formal language, with an emphasis on aggregate queries [22, 19]. The controlled experiment required to establish actual causality is approximated by a database *intervention*, i.e., by the removal of tuples that satisfy a certain description. Key challenges in this approach are 1) defining scoring functions for explanations, and 2) finding and returning the best explanations [18].

Subgroup discovery Subgroup discovery (SD) [7, 21] is concerned with finding descriptions of subsets of a dataset that have a substantial deviation in a property of interest, relative to the entire dataset; see Atzmueller [1] for a recent overview. The property of interest, or *target*, is typically an aggregate of a chosen attribute, e.g., the mean of a numeric attribute. SD algorithms typically rely on bounds on the measure of deviation to prune the search space [13].

One of the crucial shortcomings in traditional SD is the *redundancy* of results, i.e., the situation wherein the subgroups with the highest quality contain many variations of the same theme and describe only few interesting subsets. Therefore, a wide range of approaches, including the one proposed in this paper, aim at eliminating redundancy in SD. Below we briefly discuss a number of existing methods; an empirical comparison is presented in Section 6.

Sequential covering schemes, e.g., CN2-SD [8], prune or penalize subgroups that overlap with higher-ranked subgroups. Likewise, in cascaded SD [4], sub-

groups that essentially improve regression accuracy for undescribed instances are incrementally added to the result set. Although we also compare subgroups by analyzing the subsets of the data that they describe (by means of a database intervention), we do not aim at incrementally constructing a single subgroup list. Impact rules [5] and generalization-aware SD [14] prune subgroups that do not improve on their (shorter) ancestors. Unlike these methods, we also relate and compare subgroups that do not share any part of their description. Skylines of subgroup sets [11] explicate the trade-off between quality and redundancy of a set of subgroups by building the Pareto front for the given dataset and target.

3 Preliminaries

In the following, we assume the data D to consist of n rows and $m+1$ attributes. There are m *description attributes* $x_1, \ldots, x_m$, and a single *target attribute* y. The domains of x_i need not be bounded; each domain can be either categorical (nominal or ordinal) or quantitative. In the definitions we assume y to be quantitative, i.e., $y \in \mathbb{R}$, but the results can be trivially extended to the common Boolean setting, $y \in \{\text{FALSE}, \text{TRUE}\}$, by considering the proportion of TRUEs instead of the mean when computing the quality or effect size of a subgroup.

In subgroup discovery, a *subgroup description* S is usually a conjunction of conditions on the description attributes, where every condition is of the form $x_i \odot v_i$, where $\odot$ is one of $<, >, \geq, \leq, =$, and v_i is some value from the domain of x_i. For example, $S = \{x_1 \leq 0.4 \text{ AND } x_3 = 1\}$. The set of all such descriptions constitutes the pattern language $\mathcal{L}$. However, the methods we discuss in this paper are agnostic of the particular type of description language and subgroup mining algorithm being used.

The *(subgroup) cover* of S, denoted $c_D(S)$, are the rows in data D that satisfy description S. One could also think of the subgroup description as a query, and the cover as the result set of this query. As a special case, we denote by $c_D(\emptyset)$ all rows of D, i.e., an empty description matches all rows in D. The (cover) size of a subgroup is defined as the number of data rows it covers, i.e., $|c_D(S)|$.

Define the *effect* of S in data D as

$$q_D(S) = \sum_{i \in c_D(S)} y_i. \tag{1}$$

The *average effect* of S in D is then defined as

$$\mu_D(S) = \frac{q_D(S)}{|c_D(S)|}. \tag{2}$$

Also, we denote by μ_D the mean of the target attribute in the entire data D.

Given the previous, the traditional subgroup discovery task is to find the top-k subgroup descriptions with regard to some quality measure $\phi : \mathcal{L} \to \mathbb{R}$. Quality measures typically combine the size of the subgroup cover with the observed deviation in the target attribute. Probably the best known quality measure is Weighted Relative Accuracy (WRAcc), which we will formally define when we need it in Section 6.

4 Explanation Networks

Before we formalize *explanation networks*, we first describe how to adapt ideas from database causality to define to what extent subgroups "explain" each other.

4.1 Interventions as explanations

For the moment, we consider comparing two subgroups, S and T. Recent research [22, 19] in the database community has developed methods that can be used to explain away outliers (deviations) in aggregate queries in relational databases. As the basic mechanism is that of an intervention and the goal is to explain deviations, this research area is often called *database causality*. We adopt a similar technique to quantify how much the effect of subgroup S can be explained by subgroup T.

The database causality approach is based on the simple principle where individual data items are the fundamental contributing factors to all observed effects. In an *intervention*, part of the data are removed, and we observe what happens to the deviation (of some target) in the data that remains. If this deviation is substantially changed after the removal of some data, we can conclude that the removed data play an important part in the observed deviation, and thus in part *explain* the observed deviation. Given a query that exhibits anomalous behavior, the goal is to find those queries that reduce this anomalous behavior the most, the idea being that those are likely to be (causal) explanations of the deviation.

This setting is strikingly similar to the subgroup discovery setting that we consider: subgroup descriptions can be interpreted as queries and we are also interested in the deviation in some target. Let us therefore translate Wu and Madden's [22] influence definition to our notation. First, we slightly abuse notation, and let $D \setminus T$ denote $D \setminus c_D(T)$ for short, i.e., $D \setminus T$ are those rows in data D that *do not belong* to the cover of subgroup description T. Then, the *(database) influence* of a subgroup T on S is defined as

$$\text{infl}(S, T) = \frac{q_D(S) - q_{D \setminus T}(S)}{|c_D(S) \cap c_D(T)|}. \tag{3}$$

By Equation 1, $q_{D \setminus T}(S)$ is the effect of S in data where subgroup T *is not true*. Informally, we compare the effect of S in data D to the effect of S in data from which $c_D(T)$ has been removed, normalized by the number of data rows that satisfy both S and T.

Considering the difference in effect of S with and without T is a natural choice, as it is this effect that we are trying to explain. Averaging over the number of affected data rows, however, causes a strong bias towards smaller explanations: the smaller $|c_D(S) \cap c_D(T)|$, the larger the influence. This is undesirable, in particular in the subgroup discovery setting, as subgroups with small covers tend to have long descriptions and do not generalize well. In practice, this results in many subgroups consisting of very few data rows that together 'explain' the larger subgroups.

No normalization at all, on the other extreme, is not an option either: in that case subgroups T having a large cover are more likely to have a large influence. In fact, without the denominator in Equation 3 we would have $\text{infl}(S, \emptyset) = q_D(S) - q_{D \setminus D}(S) = q_D(S)$, implying that all data $D = c_D(\emptyset)$ always has the largest possible influence on any subgroup S. (Recall that $\emptyset$ is the empty subgroup description that matches all of D.)

Motivated by the previous, the solution that we propose is to compare the observed influence of T on S to the *expected influence of a random subset of data rows having (roughly) the same size as* $c_D(T)$. The aim of this is to reduce the influence of subgroups that have a large cover, as their influence would otherwise be disproportionally strong. Let T^* denote a random subset of the rows of D, so that that every row in D has an equal probability to belong to T^*, and we have $\mathbb{E}[|T^*|] = |c_D(T)|$. Notice that rather than requiring T^* to have exactly the same size as $c_D(T)$, we only constrain it to have the same size *in expectation*. This makes the resulting calculations much simpler, while achieving the same practical outcome of normalizing the influence with respect to cover size. The *expected effect* of S in data $D \setminus T^*$ is then given by

$$\mathbb{E}[q_{D \setminus T^*}(S)] = \sum_{i \in c_D(S)} \Pr[i \notin c_D(T^*)] y_i, \tag{4}$$

$$= \left(1 - \frac{|c_D(T)|}{n}\right) \sum_{i \in c_D(S)} y_i = \left(1 - \frac{|c_D(T)|}{n}\right) q_D(S). \tag{5}$$

Observe that this definition has two desirable properties: 1) it scales linearly with $q_D(S)$, meaning that it is potentially larger for subgroups that have a large deviation in the target attribute; and 2) the expected influence is smaller for larger $c_D(T)$ (and vice versa). The *explanation* of T on S, denoted $E[S, T]$, is then defined as the difference between the *expected* effect of S in $D \setminus T^*$ and the *observed* effect of S in $D \setminus T$, i.e.,

$$E[S, T] = \mathbb{E}[q_{D \setminus T^*}(S)] - q_{D \setminus T}(S) = \left(1 - \frac{|c_D(T)|}{n}\right) q_D(S) - q_{D \setminus T}(S). \tag{6}$$

This definition makes use of the desirable properties that expected effect has and has two desirable properties itself. First, it is 0 for both $c_D(T) = D$ (i.e., when $T = \emptyset$) and $c_D(T) = \emptyset$, meaning that neither the complete dataset nor the empty set is a good explanation of any subgroup S. Observe that the definition in Equation 6 allows both positive as well as negative effects: the effect of S can both increase and decrease after the intervention, and the correction for expected effect does not exclude the possibility of either direction.

4.2 Defining the network

Next we propose a novel concept that allows us to deal with the mutual relationships between multiple subgroups. While most traditional pattern mining approaches consider the pattern lattice—i.e., the search space, as defined by the

pattern language, that search procedures typically traverse as a tree—we propose a holistic perspective instead and introduce a global *network* of patterns.

Specifically, we define a *weighted directed graph* G where individual subgroups are nodes, and two nodes, S and T, are connected with a directed edge (S, T) if S can be (partially) explained by T. The weight of edge (S, T), denoted $w(S, T)$, must be proportional to the amount with which T explains S. Clearly, we will use $w(S, T) = E[S, T]$. Formally, we have the following.

Definition 1 (Explanation Network). *Given data D and a set of subgroups $\mathcal{S}$, define the* explanation network *G as $G = (V, W)$, where $V = \mathcal{S}$ and $W = \{(S, T) \mid S, T \in \mathcal{S}\}$. Each $(S, T) \in W$ has weight $w(S, T) = E[S, T]$.*

A distinguishing feature of explanation networks is that they contain information regarding both the mutual relationships between patterns 1) from the same branch of the search tree; and 2) from *different branches* of the search tree. Especially the second property is unique in that it allows to discover and exploit overlap/redundancy across completely disjoint subgroup descriptions, which is achieved by the holistic view on the covers of all subgroups in $\mathcal{S}$. As such, explanation networks can be used for many different tasks; the next section will describe how they can be used for two tasks.

5 Using Explanation Networks

Explanation networks can be used for different purposes. In this section we describe how they can be used for two different mining tasks, i.e., 1) pattern ranking, and 2) pattern set selection.

5.1 Pattern ranking

The explanations describe pairwise relationships between two subgroups, but in certain situations it may be of interest to provide a global score or ranking. To turn the pairwise relationships into a scoring method, we propose the following. Intuitively, *subgroups that are good at explaining other subgroups should have a high score*. Especially this should hold for subgroups that are good explanations of other high scoring subgroups. This can be expressed in the following recursive definition for $score(T)$:

$$score(T) \propto \sum_{S} score(S) E[S, T]. \tag{7}$$

This definition is analogous to that of PageRank [2], the well-known ranking method for web search that uses *random walks*, i.e., stochastic processes that move between a number of states, where the probability to move to any other state only depends on the current state of the walk. The PageRank score of a page is defined as its probability in the *stationary distribution*. Indeed, if this probability is high, the page must be easy to reach, and is thus of high quality.

We adapt this idea to the context of subgroups and the explanation network: if the stationary probability of a subgroup is high, then it must be easy to reach and therefore have high explanatory power. That is, we define a random walk where subgroups are states, and the transition probability from subgroup S to subgroup T is proportional to $w(S,T)$, i.e., the weight of the edge from S to T.

Formally, given the explanation network G with N nodes, we construct a random walk as follows. Let $\mathbf{A}$ denote a square matrix, the *transition probability matrix*, where element $\mathbf{A}[i,j]$ is equal to the transition probability from state i to state j, defined in two steps as follows.

1. Define the $N \times N$ matrix $\bar{\mathbf{A}}$ so that
$$\bar{\mathbf{A}}[S,T] = \begin{cases} w(S,T) & \text{if } w(S,T) > 0, \\ 0 & \text{otherwise} \end{cases}$$
for all S and T. (Here we abuse notation slightly and denote the row and column indices that correspond to subgroups S and T simply by S and T.)
2. Define the $N \times N$ matrix $\mathbf{A}$ so that, for all S and T,
$$\mathbf{A}[S,T] = \bar{\mathbf{A}}[S,T] / \sum_{S} \bar{\mathbf{A}}[S,T].$$

For a random walk to have a stationary distribution, it must be *irreducible* and *aperiodic*. In PageRank this is commonly enforced by adding a *teleportation distribution*. Let $\mathbf{b}$ denote the teleportation distribution that satisfies $\sum_i \mathbf{b}_i = 1$, where $\mathbf{b}_i$ is equal to the probability to move from any state to state i.

Finally, the score of every subgroup is given by the PageRank vector $\mathbf{s} \in \mathbb{R}^N$, i.e., the stationary distribution of the random walk, defined by [2]:

$$\mathbf{s} = \alpha \mathbf{A}^\mathsf{T} \mathbf{s} + (1-\alpha)\mathbf{b}. \tag{8}$$

If $\alpha = 1$ the teleportation distribution has no effect; in practice we usually set $\alpha = 0.7$, meaning that $\mathbf{s}$ is mainly affected by $\mathbf{A}$ (i.e., the $E[S,T]$ values). See also Subsection 6.2 for a brief empirical study of the effect α has.

In the simplest case we can use a uniform distribution for $\mathbf{b}$, i.e., we let $\mathbf{b}_S = 1/N$ for all S. The teleportation distribution can also be used to bias the resulting scores based on some other criteria: subgroups S that have higher values of $\mathbf{b}_S$ also tend to have higher scores $\mathbf{s}_S$. This idea was used to define "personalized" variants of PageRank. [2] When scoring subgroups, we can define $\mathbf{b}$ so that $\mathbf{b}_S$ is proportional to, e.g., effect size $q_D(S)$, or cover size $|c_D(S)|$. PageRank thus allows to combine different ranking criteria using the same framework.

5.2 Pattern set selection

While a pattern ranking can be of interest in a wide range of scenarios, e.g., when using patterns as input for a next analysis phase, under certain circumstances it can be more useful to have a *set of non-redundant patterns*. When patterns are

to be presented to domain experts, for example, the result set should be small. In these cases we can benefit from the explanation network by selecting *a set of patterns that explain numerous yet distinct patterns in the network*.

As with pattern ranking, we observe that this problem strongly resembles a well-known problem in network analysis: in this case, the *influence maximization (InfMax) problem* [6] in social networks. The InfMax problem concerns the selection of the k nodes in a network that are together the most influential, where influence is defined in terms of an influence *propagation model*.

We use the *Independent Cascade Model* (ICM) [6], because of its simplicity and nice theoretical properties (discussed below). The ICM assumes every node of the network to be either *active* or *inactive*. Initially all nodes are inactive, except a *seed set* of k nodes that are active. At every round, nodes that became active in the previous round (in the 1st round the seed nodes) attempt to activate their immediate neighbors. Each activation attempt is independent, and succeeds with probability $\mathbf{P}[T, S]$, where T is an active node and S is an inactive node. The process finishes when no activation attempt in a round is successful. The *influence of the seed set* is *the number of active nodes when the process finishes*.

To adapt this idea for subgroup set selection, we solve the InfMax problem on the explanation network G with appropriately defined activation probabilities $\mathbf{P}[T, S]$. Intuitively, as we want to find a pattern set that has a high explanation strength, we let $\mathbf{P}[T, S] \propto w(S, T)$. I.e., T activates S with a probability that is *directly proportional to the explanation of subgroup T on subgroup S*. In practice we let $\mathbf{P}[T, S] = \mathbf{A}[S, T]$, where $\mathbf{A}[S, T]$ is defined in the same way in Section 5.1. However, our approach is by no means tied to ICM; any propagation model that can be parametrized in terms of the network weights $w(S, T)$ can be used.

The *explanation maximization problem* is then defined as *finding those k subgroups that maximize influence when chosen as the seed set*. Kempe et al. [6] showed that solving the InfMax problem is in general NP-hard, but also that the ICM results in a *submodular* influence function. Therefore the problem can be solved efficiently by a greedy algorithm that one at a time selects the node that maximizes marginal gain in influence. This algorithm has a constant approximation ratio, i.e., it provides a solution of size k that has influence at least $(1 - 1/e)$ times the influence of the optimal solution. All experiments in this paper are carried out using this algorithm.

Finally, an important aspect that differentiates the explanation network from, e.g., social networks is its density, i.e., it contains substantially more edges in relation to the number of vertices. In practice the explanation network can be a single, large clique. As the complexity of seed selection algorithms mostly depends on the number of edges, it is very important to use an efficient algorithm despite the number of edges often being much lower than in social networks.

6 Experiments

In this section we evaluate how well the two tasks based on explanation networks that we introduced perform and empirically compare them to existing methods.

Baseline methods: We consider the following four baseline methods, as they are well-known and the latter two are representative of the state of the art:

NRAcc: Rank patterns in decreasing order of Normalized Relative Accuracy, i.e., $(\mu_D(S)-\mu)/\sigma$, where μ and σ denote global mean and standard deviation of the target attribute.

WRAcc: Rank patterns in decreasing order of Weighted Relative Accuracy, defined as $\sqrt{|S|} \times$ NRAcc.

Generalization aware pruning (gap): Rank patterns in decreasing order of the *gap* score, defined as $\mu_D(S) - \max_{S'} \mu_D(S')$, where S' is a subgroup that is a *generalization* of S (i.e., its description consists of a subset of the conditions in the description of S).

Greedy-WRAcc: Sequential covering using WRAcc: iteratively select that subgroup that maximizes WRAcc (as defined above), remove the rows that belong to its cover from the data, and iterate until enough subgroups have been selected or until the data is exhausted.

Real data: As datasets we use the Abalone (aba), Credit-G (cg), Mushroom (mush), Redwine (rw), and Wages (wag) from the UCI Machine Learning repository[4]. Further, we also include Elections (ele) as described in [10] and the Helsinki housing (hel) [23] dataset. cg and mush have a Boolean target, all others have a numeric target. On-the-fly discretisation of numeric description attributes was applied, meaning that 6 equal-size intervals were created upon pattern extension. Dataset sizes range up to 8337 rows (for hel) and 73 attributes (for ele).
Subgroup candidates: In the experiments we assume a fixed candidate set of subgroups. The candidate set is obtained by mining subgroups using NRAcc (as defined above) using a support threshold of 10% and maximum search depth of 3. If there were more than 5000 candidates, these were initially ranked in terms of WRAcc, and the top-5000 were kept for the experiment.
Note on running times / complexity: Constructing the explanation network requires computing intersections of $c_D(S)$ between all pairs of subgoups. Pagerank is computed using the basic power-iteration method, which converges rapidly in practice. For the seed selection task we use the greedy algorithm and some simple optimizations to speed up influence computations. Our methods[5] complete in approximately 15 minutes for the largest datasets (hel, elections), and in less than a minute for the smaller ones.

6.1 Artificial data generation

We employ two approaches for creating artificial data with planted subgroups. The first one is based on a Bayesian network model, while the second one combines a real dataset with an artificial target attribute.

Bayesian network This generative model (illustrated in Figure 1) consists of a number of independent *causal chains* with variables X that all end in the

[4] http://archive.ics.uci.edu/ml/
[5] Source code is available at: http://anttiukkonen.com/explanation-networks/

target attribute Y, and a number of random attributes R that are independent of everything, including Y. The idea is to model several different causes for observing $Y = 1$ in the underlying process. Ideally we find subgroups where the description does not contain any random attributes.

Here X_1^i is the *root cause* of the output Y in chain i, the other X_j^is are intermediary effects. The conditional probabilities of the X_j^i variables are adjusted so that $X_j^i = 1$ almost always when $X_{j-1}^i = 1$, and $X_j^i = 0$ almost always when $X_{j-1}^i = 0$. Also, $Y^i = 1$ almost always when $X_h^i = 1$, and $Y^i = 0$ almost always when $X_h^i = 0$. Finally, $Y = 1$ whenever at least one $Y^i = 1$, and $Y = 0$ otherwise. Data from the model is generated by first drawing a dataset of size N separately from every chain. These are combined by concatenating the vectors and by introducing the global target variable Y that is simply the union of the Y^is from every chain. The original Y^is are removed. To generate data from one of the causal chains, we compute the exact joint distribution of the X_j^i and Y^i variables, and then draw N binary vectors from this distribution. In addition, we add q "non-causal" noise variables $R_1^i, \ldots, R_q^i$ by creating randomly permuted copies of some of the X_j^i.

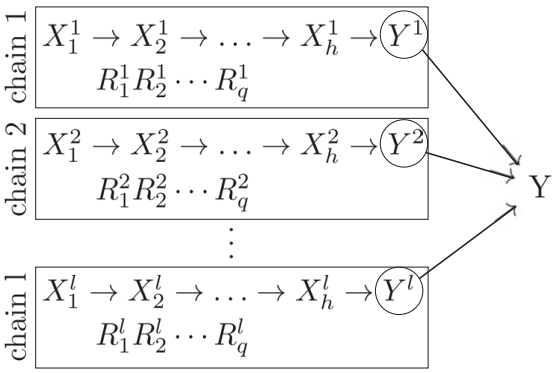

Fig. 1. A Bayesian network with l independent "chains", and a single target attribute Y.

Below we refer to data sampled from this model using the notation BN(h, q, l). For example, BN(2,4,5) refers to data from a model with five chains, each containing two causal variables and four noise variables (that is, 30 variables in total that serve as description attributes, plus one target).

Latent cause model In a manner similar to the approach described above, we aim to simulate a scenario where the objective is to uncover the "true" cause of a phenomenon. In the Bayes network model this true cause was expressed by the observed X_j^i variables. Now we increase the level of difficulty and assume that the true cause is unobserved: it is only reflected in a noisy, numeric target attribute. Moreover, we assume that no perfect description of the true cause exists. In concrete, the true underlying cause is an unobserved binary attribute Z. The observed target Y is generated from Z by drawing a random "low" value for those rows for which $Z = 0$, and a random "high" value for rows where $Z = 1$. The aim is to find subgroups the covers of which are a good match with $Z = 1$ (and do not necessarily have a high quality w.r.t. the observed target).

To maintain a realistic structure of the search space, we start from a given real dataset D and replace its original target attribute with an artificial one:

1. Select a random subset Z of the rows of D, s.t. $|Z| = 1/3|D|$. This corresponds to the "true" underlying cause.

2. Find a set of random subgroups. Select K of these s.t. their covers match the set Z as well as possible. (We used the F1-measure to calculate the goodness of the match.) These are the subgroups that we aim to find.
3. Create a target attribute Y where the value for rows in Z is drawn from a normal distribution having mean 2 and stdev 1, while the value for the other rows is drawn from a normal distribution with mean 0 and stdev 1. Replace the original target of D with Y.

6.2 Pattern ranking

In the first experiment we use data from the Bayes network model to study what effect the α parameter of Eq. 8 has on the resulting pattern ranking. We define the teleportation distribution using the normalized relative accuracies of the subgroups, i.e., we let $\mathbf{b}_S \propto \frac{\mu_D(S)-\mu}{\sigma}$ for every S, where μ and σ are the mean and standard deviation, respectively, of the target attribute in D. Now α can be understood as a parameter that adjusts the effect between NRAcc and explanation strength of a subgroup. (For $\alpha = 0$ the ranking is only based on NRAcc, while for $\alpha = 1$ it is only based on the explanation strengths.)

With Bayes network data we can evaluate performance in terms of AUC by treating the pattern ranking problem as a classification problem where the objective is to separate those subgroups that have "non-causal" attributes in their description from those that only have "causal" attributes. We consider all subgroups that do not have *any noise* variables R_i in their descriptions as "causal". That is, we are very strict and want to find such subgroups that *only* describe phenomena that are associated with the target according to our model.

Results are shown in Fig. 2. From the two top-most panels we can observe that the PageRank-based pattern ranking approach performs well in terms of AUC. The lines are average AUCs over 20 independently drawn datasets (of 5000 rows) from the respective models, and the dashed lines show naïve confidence bands of $\pm$ 3 standard deviations. Especially when the target has several independent causes (BN(5,5,5)), we find the explanation network to show significant improvements over using WRAcc (or NRAcc) only. The panels in the bottom row of Fig. 2 show how explanation based pagerank is related to both WRAcc and subgroup size in subgroups mined from BN(5,5,5) (5000 rows) with $\alpha = 0.85$. Indeed, we can observe that pagerank separates the "causal" subgroups (shown in blue) from the "non-causal" subgroups (red) better than WRAcc. Moreover, while the explanation based ranking tends to favor large subgroups, simply ranking subgroups by size would not give the same result either.

6.3 Pattern set selection

We continue with an experiment where the task is always to *retrieve a pattern set of 20 subgroups* from the given candidate set. Results with Bayes network data are evaluated in terms of *precision of retrieving "causal" subgroups*, i.e., the fraction of such subgroups in the 20 subgroups returned. Results with latent

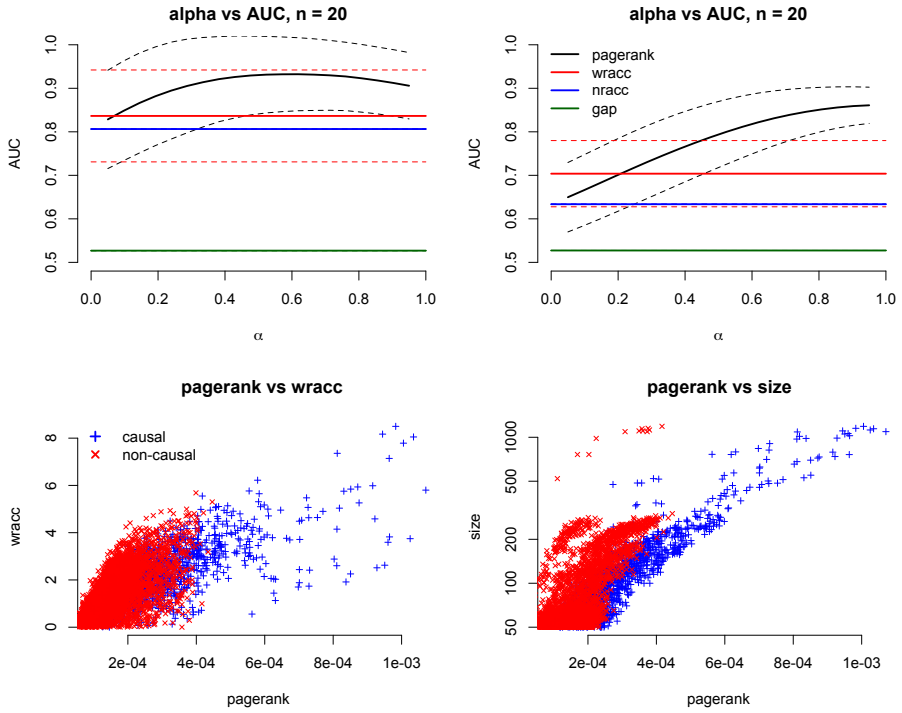

Fig. 2. Top-left: Mean AUC (solid) and naïve confidence bands (dashed) as a function of α using BN(3,6,3) data (see legend of top-right). Top-right: AUC as a function of α using BN(5,5,5) data. Bottom-left: Pagerank score vs. WRAcc for subgroups from BN(5,5,5). Bottom-right: Pagerank score vs. size for subgroups from BN(5,5,5).

cause model are evaluated in terms of *precision of retrieving "correct" subgroups*, where a subgroup is considered as "correct" if it was chosen in step 2 of the target generation procedure ($K = 20$). Finally, results with real, unmodified data are evaluated in terms of a score function that is composed of four quantities: 1) average cover size (*avg.size*), 2) average quality (*avg.qual)*, 3) entropy of the cover distribution (*cent*), and 4) fraction of data rows that are covered by at least one of the chosen subgroups (*ccov*). These are computed for all methods, and normalized to $[0, 1]$ by dividing with the value attained by the best performing method. The final score, denoted PSS (for "pattern set score"), is the *geometric mean* of these normalized numbers. We use the geometric mean because all four quantities are important and poor performance in even only one of them is undesirable. This score is also shown for the artificial datasets.

Results with Bayes network data are shown in Fig. 3, where we plot precision against pattern set score for all methods under different parametrizations of the Bayes network. Our methods (P and S, shown in red) have both higher precision as well as PSS score for all but the most simplest model, BN(2,2,2).

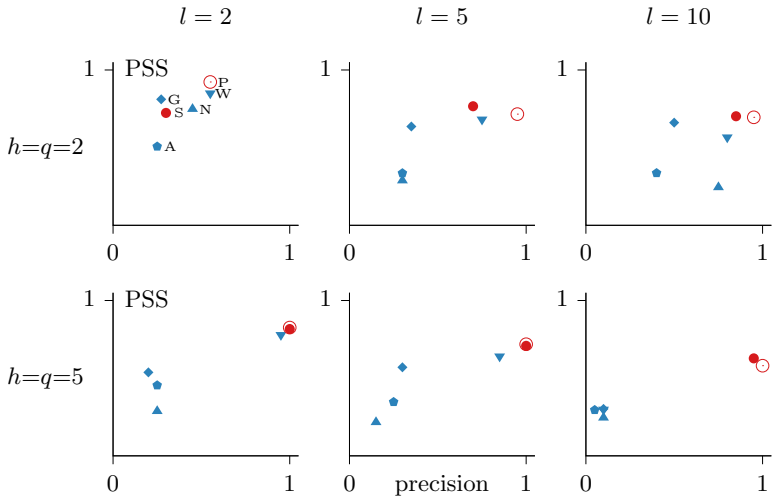

Fig. 3. Precision and Pattern Set Scores for experiments with Bayes data. The proposed explanation based methods "pagerank" (P) and "seeds" (S) outperform the competitors in the hard settings with larger numbers of antecedents (h) and component chains (l).

Results with latent cause data are shown in Table 1. This is a very hard task, as there are very few planted subgroups, and these are by definition not very well correlated with the noisy target. The explanation based approaches have the highest aggregate scores. Furthermore, they are the only methods that succeed in discovering some of the planted subgroups.

Finally, we present average evaluation metric values over the real datasets in Table 2. We find that both algorithms that are based on a greedy selection heuristic, seeds and greedy-wracc, have the same average score. However, this score is composed differently for the two methods. Greedy-wracc has a higher entropy (cent) and cover (ccov) value, while seeds performs better in terms of average cover size and subgroup quality. Overall, evaluating subgroup sets in an objective, application independent manner, is difficult, and it is not obvious that the same method is appropriate for all tasks. However, the results of Ta-

Table 1. Experiments with latent cause data (averages over 20 randomized runs).

	PSS				precision			
	aba	cg	mush	wag	aba	cg	mush	wag
pagerank	**0.54**	**0.43**	**0.48**	**0.51**	0.00	**0.25**	**0.25**	0.00
seeds	**0.58**	**0.52**	**0.53**	**0.52**	0.00	**0.30**	**0.40**	**0.05**
gap	0.46	0.30	0.29	0.41	0.00	0.00	0.00	0.00
greedy-wracc	0.45	**0.39**	0.36	**0.50**	0.00	0.00	0.00	0.00
nracc	0.23	0.16	0.19	0.13	0.00	0.00	0.00	0.00
wracc	0.40	0.23	0.28	0.33	0.00	0.00	0.00	0.00

Table 2. Experiments with real data (averages over all seven real datasets).

	PSS	avg.size	ccov	cent	avg.qual
pagerank	0.56	**0.84**	0.49	0.33	0.76
seeds	**0.74**	**0.93**	**0.78**	0.67	0.66
gap	**0.68**	0.58	**0.75**	**0.77**	0.73
greedy-wracc	**0.74**	0.65	**0.98**	**0.86**	0.61
nracc	0.49	0.37	0.38	0.47	**1.00**
wracc	0.59	0.74	0.51	0.44	**0.82**

ble 2 suggest that the explanation based approaches (pagerank and seeds) find subgroups of reasonably high quality that have a fairly large cover, meaning the subgroups should generalize better to unseen data.

7 Conclusions

We introduced *explanation networks*, a novel, global perspective on subgroups and how they relate to each other. In particular, we used interventions to define the notion of explanation, which quantifies the explanatory influence of one subgroup on another and is normalized by the expected influence from a random subgroup of the same size. We showed how analogies with network analysis can be made and how they can lead to novel pattern mining methods. In this paper we have studied the use of PageRank for pattern ranking and the use of seed selection (influence maximization) for pattern set selection.

The experiments demonstrate that our explanation based approach provides advantages when it is of importance to select subgroup descriptions that capture the data generating process. Specifically, on artificial data we have shown—using very strict evaluation criteria—that our approach provides better rankings and pattern sets than the competitors.

Although these results clearly show the potential of explanation networks, there are also still many directions to be explored. For example, we will perform user studies in which the analyst is enabled to visually explore the network for alternative explanations. Further, it is of interest to investigate direct exploration and mining algorithms that avoid the need to materialise the full explanation network. As a third and final example, it would be interesting to develop a statistical test for assessing whether the influence of one subgroup on another is significant. In general, much is still to be gained from this novel network perspective on explanation and redundancy in pattern mining.

Acknowledgements Antti Ukkonen was partially supported by Tekes (project Re:Know2) and Academy of Finland (decision 288814).

References

1. Atzmueller, M.: Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(1), 35–49 (2015)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks 30(1-7), 107–117 (1998)
3. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by aggregating emerging patterns. In: Proceedings of DS'99. pp. 30–42 (1999)
4. Grosskreutz, H.: Cascaded subgroups discovery with an application to regression. In: Proceedings of LeGo ECML/PKDD Workshop (2008)
5. Huang, S., Webb, G.I.: Discarding insignificant rules during impact rule discovery in large, dense databases. In: Proceedings of SDM. pp. 541–545 (2005)
6. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD. pp. 137–146 (2003)
7. Klösgen, W.: Advances in Knowledge Discovery and Data Mining, chap. Explora: A Multipattern and Multistrategy Discovery Assistant, pp. 249–271 (1996)
8. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5(Feb), 153–188 (2004)
9. Lavrač, N., Gamberger, D.: Relevancy in Constraint-Based Subgroup Discovery. In: European Workshop on Inductive Databases & Constraint Based Mining (2006)
10. van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery 25(2), 208–242 (2012)
11. van Leeuwen, M., Ukkonen, A.: Discovering skylines of subgroup sets. In: Proceedings of ECML/PKDD. pp. 273–287 (2013)
12. Leman, D., Feelders, A., Knobbe, A.J.: Exceptional Model Mining. In: Proceedings of ECML/PKDD. pp. 1–16 (2008)
13. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast exhaustive subgroup discovery with numerical target concepts. Data Mining and Knowledge Discovery 30(3), 711–762 (2016)
14. Lemmerich, F., Becker, M., Puppe, F.: Difference-based estimates for generalization-aware subgroup discovery. In: Proceedings of ECML/PKDD. pp. 288–303 (2013)
15. Meliou, A., Gatterbauer, W., Halpern, J.Y., Koch, C., Moore, K.F., Suciu, D.: Causality in databases. IEEE Data Eng. Bull. 33(3), 59–67 (2010)
16. Meliou, A., Roy, S., Suciu, D.: Causality and explanations in databases. Proceedings of the VLDB Endowment 7(13), 1715–1716 (2014)
17. Pearl, J.: Causality. Cambridge University Press, 2nd edn. (2009)
18. Roy, S., Orr, L., Suciu, D.: Explaining query answers with explanation-ready databases. Proceedings of the VLDB Endowment 9(4), 348–359 (2015)
19. Roy, S., Suciu, D.: A formal approach to finding explanations for database queries. In: Proceedings of SIGMOD. pp. 1579–1590 (2014)
20. Terada, A., Okada-Hatakeyama, M., Tsuda, K., Sese, J.: Statistical significance of combinatorial regulations. Proceedings of the National Academy of Sciences 110(32), 12996–13001 (2013)
21. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proceedings of PKDD. pp. 78–87 (1997)
22. Wu, E., Madden, S.: Scorpion: Explaining away outliers in aggregate queries. Proceedings of the VLDB Endowment 6(8), 553–564 (2013)
23. Zliobaite, I., Mathioudakis, M., Lehtiniemi, T., Parviainen, P., Janhunen, T.: Accessibility by public transport predicts residential real estate prices: A case study in helsinki region. In: 2nd Workshop on Mining Urban Data at ICML 2015 (2015)