DEvIANT: Discovering Significant Exceptional (Dis-)Agreement Within Groups

Adnene Belfodil, Wouter Duivesteijn, Marc Plantevit, Sylvie Cazalens, Philippe Lamarre

DEvIANT: Discovering Significant Exceptional (Dis-)Agreement Within Groups [Tech. Report]

Adnene Belfodil[1] (✉), Wouter Duivesteijn[2], Marc Plantevit[3], Sylvie Cazalens[1], and Philippe Lamarre[1]

[1] Univ Lyon, INSA Lyon, CNRS, LIRIS UMR 5205, F-69621, Lyon, France
[2] Technische Universiteit Eindhoven, Eindhoven, the Netherlands
[3] Univ Lyon, CNRS, LIRIS UMR 5205, F-69622, Lyon, France

Abstract. We strive to find contexts (i.e., subgroups of entities) under which exceptional (dis-)agreement occurs among a group of individuals, in any type of data featuring individuals (e.g., parliamentarians, customers) performing observable actions (e.g., votes, ratings) on entities (e.g., legislative procedures, movies). To this end, we introduce the problem of discovering statistically significant exceptional contextual intra-group agreement patterns. To handle the sparsity inherent to voting and rating data, we use Krippendorff's Alpha measure for assessing the agreement among individuals. We devise a branch-and-bound algorithm, named DEvIANT, to discover such patterns. DEvIANT exploits both closure operators and tight optimistic estimates. We derive analytic approximations for the confidence intervals (CIs) associated with patterns for a computationally efficient significance assessment. We prove that these approximate CIs are nested along specialization of patterns. This allows to incorporate pruning properties in DEvIANT to quickly discard non-significant patterns. Empirical study on several datasets demonstrates the efficiency and the usefulness of DEvIANT.

1 Introduction

Consider data describing voting behavior in the European Parliament (EP). Such a dataset records the votes of each member (MEP) in voting sessions held in the parliament, as well as the information on the parliamentarians (e.g., gender, national party, European party alliance) and the sessions (e.g., topic, date). This dataset offers opportunities to study the agreement or disagreement of coherent subgroups, especially to highlight unexpected behavior. It is to be expected that on the majority of voting sessions, MEPs will vote along the lines of their European party alliance. However, when matters are of interest to a specific nation within Europe, alignments may change and agreements can be formed or dissolved. For instance, when a legislative procedure on fishing rights is put before the MEPs, the island nation of the UK can be expected to agree on a specific course of action regardless of their party alliance, fostering an exceptional agreement where strong polarization exists otherwise.

We aim to discover such exceptional (dis-)agreements. This is not limited to just EP or voting data: members of the US congress also vote on bills, while

Amazon-like customers post ratings or reviews of products. A challenge when considering such voting or rating data is to effectively handle the absence of outcomes (sparsity), which is inherently high. For instance, in the European parliament data, MEPs vote on average on only ¾ of all sessions. These outcomes are not missing at random: special workgroups are often formed of MEPs tasked with studying a specific topic, and members of these workgroups are more likely to vote on their topic of expertise. Hence, present values are likely associated with more pressing votes, which means that missing values need to be treated carefully. This problem becomes much worse when looking at Amazon or Yelp rating data: the vast majority of customers will not have rated the vast majority of products/places.

We introduce the problem of discovering significantly exceptional contextual intra-group agreement patterns, rooted in the Subgroup Discovey (SD) [47]/ Exceptional Model Mining (EMM) [8] framework. To tackle the data sparsity issue, we measure the agreement among groups with *Krippendorff's alpha*, a measure developed in the context of content analysis [28] which handles missing outcomes elegantly. We develop a branch-and-bound algorithm to find subgroups featuring statistically significantly exceptional (dis-)agreement among groups. This algorithm enables discarding non-significant subgroups by pruning unpromising branches of the search space (cf. Figure 1). Suppose that we are interested in subgroups of entities (e.g., voting sessions) whose sizes are greater than a support threshold σ. We gauge the exceptionality of a given subgroup of size $X \geq \sigma$, by its *p-value*: the probability that for a random subset of entities, we observe an intra-agreement at least as extreme as the one observed for the subgroup.

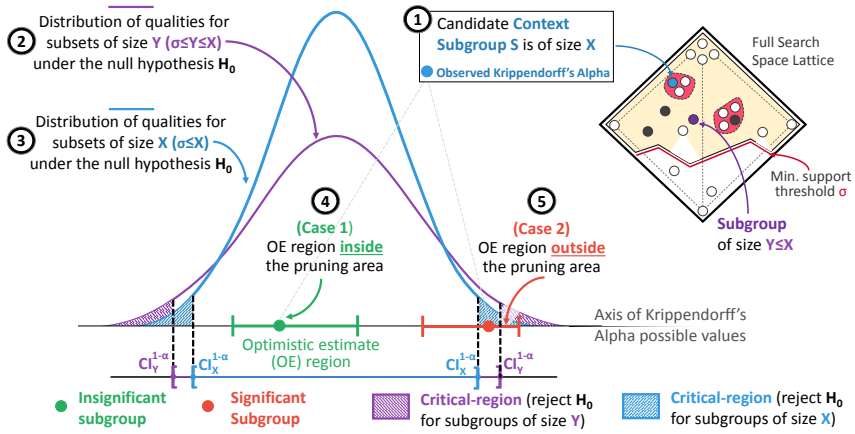


Fig. 1: Main DEvIANT properties for safe sub-search space pruning. A subgroup is reported as significant if its related Krippendorff's Alpha falls in the critical region of the corresponding empirical distribution of random subsets (DFD). When traversing the search space downward (decreasing support size), the approximate confidence intervals are nested. If the optimistic estimates region falls into the confidence interval computed on the related DFD, the sub-search space can be safely pruned.

Thus we avoid reporting subgroups observing a low/high intra-agreement due to chance only. To achieve this, we estimate the empirical distribution of the intra-agreement of random subsets (DFD: Distribution of False Discoveries, cf. [9,33]) and establish, for a chosen critical value α, a confidence interval $CI_X^{1-\alpha}$ over the corresponding distribution under the null hypothesis. If the subgroup intra-agreement is outside $CI_X^{1-\alpha}$, the subgroup is statistically significant ($p\text{-}value \leq \alpha$); otherwise the subgroup is a spurious finding. We prove that the analytic approximate confidence intervals are nested: $\sigma \leq Y \leq X \Rightarrow CI_X^{1-\alpha} \subseteq CI_Y^{1-\alpha}$ (i.e., when the support size grows, the confidence interval shrinks). Moreover, we compute a tight optimistic estimate (OE) [18] to define a lower and upper bounds of Krippendorff's Alpha for any specialization of a subgroup having its size greater than σ. Combining these properties, if the OE region falls into the corresponding CI, we can safely prune large parts of the search space that do not contain significant subgroups. In summary, the main contributions are:
1) We introduce the problem of discovering statistically significant exceptional contextual intra-group agreement patterns (Section 3).
2) We derive an analytical approximation of the confidence intervals associated with subgroups. This allows a computationally efficient assessment of the statistical significance of the findings. Furthermore, we show that approximate confidence intervals are nested (Section 4). Particular attention is also paid to the variability of outcomes among raters (Section 5).
3) We devise a branch-and-bound algorithm to discover exceptional contextual intra-group agreement patterns (Section 6). It exploits tight optimistic estimates on Krippendorff's alpha and the nesting property of approximate CIs.

2 Background and Related Work

The page limit, combined with the sheer volume of other material in this paper, compels us to restrict this section to one page containing only the most relevant research to this present work.

Measuring Agreement. Several measures of agreement focus on two targets (Pearson's ρ, Spearman's ρ, Kendall's τ, Association); most cannot handle missing values well. As pointed out by Krippendorff [28, p.244], using association and correlation measures to assess agreement leads to particularly misleading conclusions: when all data falls along a line $Y = aX + b$, correlation is perfect, but agreement requires that $Y = X$. Cohen's κ is a seminal measure of agreement between two raters who classify items into a fixed number of mutually exclusive categories. Fleiss' κ extends this notion to multiple raters and requires that each item receives the exact same number of ratings. Krippendorff's alpha generalizes these measures while handling multiple raters, missing outcomes and several metrics [28, p.232].
Discovering Significant Patterns. Statistical assessment of patterns has received attention for a decade [44,21], especially for association rules [20,35]. Some work focused on statistical significance of results in SD/EMM during enumeration [9,33] or a posteriori [10] for statistical validation of the found subgroups.

Voting and Rating Data Analysis. Previous work [3] proposed a method to discover exceptional *inter*-group agreement in voting or rating data. This method does not allow to discover *intra*-group agreement. In rating datasets, groups are uncovered whose members exhibit an agreement or discord [6] or a specific rating distribution [1] (e.g., polarized, homogeneous) given upfront by the end-user. This is done by aggregating the ratings through an arithmetic mean or a rating distribution. However, these methods do not allow to discover exceptional (dis-)agreement within groups. Moreover, they may output misleading hypotheses over the intra-group agreement, since aggregating ratings in a distribution (i) is highly affected by data sparsity (e.g., two reviewers may significantly differ in their number of expressed ratings) and (ii) may conceal the true nature of the underlying intra-group agreement. For instance, a rating distribution computed for a collection of movies may highlight a polarized distribution of ratings (interpreted as a disagreement) while ratings over each movie may describe a consensus between raters (movies are either highly or lowly rated or by the majority of the group). These two issues are addressed by Krippendorff's alpha.

3 Problem Definition

Our data consists of a set of individuals (*e.g., social network users, parliamentarians*) who give outcomes (*e.g., ratings, votes*) on entities (*e.g., movies, ballots*). We call this type of data a *behavioral dataset* (cf. Table 1).

Definition 1 (*Behavioral Dataset*). *A behavioral dataset* $\mathcal{B} = \langle G_I, G_E, O, o \rangle$ *is defined by (i) a finite collection of Individuals* G_I*, (ii) a finite collection of Entities* G_E*, (iii) a domain of possible Outcomes* O*, and (iv) a function* $o : G_I \times G_E \rightarrow O$ *that gives the outcome of an individual* i *over an entity* e*.*

The elements from G_I (resp. G_E) are augmented with descriptive attributes $\mathcal{A}_I$ (resp. $\mathcal{A}_E$). Attributes $a \in \mathcal{A}_I$ (resp. $\mathcal{A}_E$) may be Boolean, numerical or categorical, potentially organized in a taxonomy. Subgroups (subsets) of G_I (resp. G_E) are defined using descriptions from $\mathcal{D}_I$ (resp. $\mathcal{D}_E$). These descriptions are formalized by conjunctions of conditions on the values of the attributes. Descriptions of $\mathcal{D}_I$ are called *groups*, denoted g. Descriptions of $\mathcal{D}_E$ are called *contexts*,

Table 1: Example of behavioral dataset - European Parliament Voting dataset

(a) Entities

ide	themes	date
e_1	1.20 Citizen's rights	20/04/16
e_2	5.05 Economic growth	16/05/16
e_3	1.20 Citizen's rights; 7.30 Judicial Coop	04/06/16
e_4	7 Security and Justice	11/06/16
e_5	7.30 Judicial Coop	03/07/16
e_6	7.30 Judicial Coop	29/07/16

(b) Individuals

idi	country	group	age
i_1	France	S&D	26
i_2	France	PPE	30
i_3	Germany	S&D	40
i_4	Germany	ALDE	45

(c) Outcomes

idi	ide	o(i,e)	idi	ide	o(i,e)
i_1	e_2	Against	i_3	e_1	For
i_1	e_5	For	i_3	e_2	Against
i_1	e_6	Against	i_3	e_3	For
i_2	e_1	For	i_3	e_5	Against
i_2	e_3	Against	i_4	e_1	For
i_2	e_4	For	i_4	e_4	For
i_2	e_5	For	i_4	e_6	Against

denoted c. From now on, G (resp. $\mathcal{D}$) denotes both collections G_I (resp. $\mathcal{D}_I$) and G_E (resp. $\mathcal{D}_E$) if no confusion can arise. We denote by G^d the subset of records characterized by the description $d \in \mathcal{D}$. Descriptions from $\mathcal{D}$ are partially ordered by a specialization operator denoted $\sqsubseteq$. A description d_2 is a specialization of d_1, denoted $d_1 \sqsubseteq d_2$, if and only if $d_2 \Rightarrow d_1$ from a logical point of view. It follows that $G^{d_2} \subseteq G^{d_1}$.

3.1 Intra-group Agreement Measure: Krippendorff's Alpha (A)

Krippendorff's Alpha (denoted A) measures the agreement among raters. This measure has several properties that make it attractive in our setting, namely: (i) it is applicable to any number of observers; (ii) it handles various domains of outcomes (ordinal, numerical, categorical, time series); (iii) it handles missing values; (iv) it corrects for the agreement expected by chance. A is defined as:

$$A = 1 - \frac{D_{\text{obs}}}{D_{\text{exp}}} \tag{1}$$

where D_{obs} (resp. D_{exp}) is a measure of the observed (resp. expected) disagreement. Hence, when $A = 1$, the agreement is as large as it can possibly be (given the class prior), and when $A = 0$, the agreement is indistinguishable to agreement by chance. We can also have $A < 0$, where disagreement is larger than expected by chance and which corresponds to systematic disagreement.

Given a behavioral dataset $\mathcal{B}$, we want to measure Krippendorff's alpha for a given context $c \in \mathcal{D}_E$ characterizing a subset of entities $G_E^c \subseteq G_E$, which indicates to what extent the individuals who comprise some selected group are in agreement $g \in \mathcal{D}_I$. From Equation (1), we have: $A(S) = 1 - \frac{D_{\text{obs}}(S)}{D_{\text{exp}}}$ for any $S \subseteq G_E$. Note that the measure only considers entities having at least two outcomes; we assume the entities not fulfilling this requirement to be removed upfront by a preprocessing phase. We capture observed disagreement by:

$$D_{\text{obs}}(S) = \frac{1}{\sum_{e \in S} m_e} \sum_{o_1 o_2 \in O^2} \delta_{o_1 o_2} \cdot \sum_{e \in S} \frac{m_e^{o_1} \cdot m_e^{o_2}}{m_e - 1} \tag{2}$$

Where m_e is the number of expressed outcomes for the entity e and $m_e^{o_1}$ (resp. $m_e^{o_2}$) represents the number of outcomes equal to o_1 (resp. o_2) expressed for the entity e. $\delta_{o_1 o_2}$ is a distance measure between outcomes, which can be defined according to the domain of the outcomes (e.g., $\delta_{o_1 o_2}$ can correspond to the Iverson bracket indicator function $[o_1 \neq o_2]$ for categorical outcomes or distance between ordinal values for ratings. Choices for the distance measure are discussed in [28]). The disagreement expected by chance is captured by:

$$D_{\text{exp}} = \frac{1}{m \cdot (m-1)} \sum_{o_1, o_2 \in O^2} \delta_{o_1 o_2} \cdot m^{o_1} \cdot m^{o_2} \tag{3}$$

Where m is the number of all expressed outcomes, m^{o_1} (resp. m^{o_2}) is the number of expressed outcomes equal to o_1 (resp. o_2) observed in the entire behavioral dataset. This corresponds to the disagreement by chance observed on the overall marginal distribution of outcomes.

Example: Table 2 summarizes the behavioral data from Table 1. The disagreement expected by chance equals (given: $m^F = 8$, $m^A = 6$): $D_{\text{exp}} = 48/91$. To evaluate intra-agreement among the four individuals in the global context (considering all entities), first we need to compute the observed disagreement $D_{\text{obs}}(G_E)$. This equals the weighted average of the two last lines by considering the quantities m_e as the weights: $D_{\text{obs}}(G_E) = \frac{4}{14}$. Hence, for the global context, $A(G_E) = 0.46$. Now, consider the context $c = \langle \textit{themes} \supseteq \{7.30 \text{ Judicial Coop.}\}\rangle$, having as support: $G_E^c = \{e_3, e_5, e_6\}$. The observed disagreement is obtained by computing the weighted average, only considering the entities belonging to the context: $D_{\text{obs}}(G_E^c) = \frac{4}{7}$. Hence, the contextual intra-agreement is: $A(G_E^c) = -0.08$.

Table 2: Summarized Behavioral Data; $D_{\text{obs}}(e) = \sum_{o_1,o_2 \in O^2} \delta_{o_1 o_2} \frac{m_e^{o_1} \cdot m_e^{o_2}}{m_e \cdot (m_e - 1)}$

	[**F**]or		[**A**]gainst			
	e_1	e_2	e_3	e_4	e_5	e_6
i_1		A			F	A
i_2	F		A	F	F	
i_3	F	A	F		A	
i_4	F			F		A
m_e	3	2	2	2	3	2
$D_{\text{obs}}(e)$	0	0	1	0	$\frac{2}{3}$	0

Comparing $A(G_E^c)$ and $A(G_E)$ leads to the following statement: *"while parliamentarians are slightly in agreement in overall terms, matters of judicial cooperation create systematic disagreement among them".*

3.2 Mining Significant Patterns with Krippendorff's Alpha

We are interested in finding patterns of the form $(g, c) \in \mathcal{P}$ (with $\mathcal{P} = \mathcal{D}_I \times \mathcal{D}_E$), highlighting an exceptional intra-agreement between members of a group of individuals g over a context c. We formalize this problem using the well-established framework of SD/EMM [8], while giving particular attention to the statistical significance and soundness of the discovered patterns [21].

Given a group of individuals $g \in \mathcal{D}_I$, we strive to find contexts $c \in \mathcal{D}_E$ where the observed intra-agreement, denoted $A^g(G_E^c)$, *significantly* differs from the expected intra-agreement occurring due to chance alone. In the spirit of [9,33,44], we evaluate pattern interestingness by statistical significance of the contextual intra-agreement: we estimate the probability to observe the intra-agreement $A^g(G_E^c)$ or a more extreme value, which corresponds to the *p-value* for some null hypothesis H_0. The pattern is said to be *significant* if the estimated probability is low enough (i.e., under some critical value α). The relevant null hypothesis H_0 is: the observed intra-agreement is generated by the distribution of intra-agreements observed on a bag of i.i.d. random subsets drawn from the entire collection of entities (DFD: Distributions of False Discoveries, cf. [9]).

Problem Statement. (*Discovering Exceptional Contextual Intra-group Agreement Patterns*). Given a behavioral dataset $\mathcal{B} = \langle G_I, G_E, O, o\rangle$, a minimum group support threshold σ_I, a minimum context support threshold σ_E, a significance critical value $\alpha \in]0, 1]$, and the null hypothesis H_0 (the observed intra-agreement is generated by the DFD); find the pattern set $P \subseteq \mathcal{P}$ such that:

$P = \{(g, c) \in \mathcal{D}_I \times \mathcal{D}_E \ : \ |G_I^g| \geq \sigma_I \text{ and } |G_E^c| \geq \sigma_E \text{ and } \textit{p-value}^g(c) \leq \alpha\}$

where $\textit{p-value}^g(c)$ is the probability (under H_0) of obtaining an intra-agreement A at least as extreme as $A^g(G_E^c)$, the one observed over the current context.

4 Exceptional Contexts: Evaluation and Pruning

From now on we omit the exponent g if no confusion can arise, while keeping in mind a selected group of individuals $g \in \mathcal{D}_I$ related to a subset $G_I^g \subseteq G_I$.

To evaluate the extent to which our findings are exceptional, we follow the significant pattern mining paradigm[4]: we consider each context c as a hypothesis test which returns a *p-value*. The *p-value* is the probability of obtaining an intra-agreement at least as extreme as the one observed over the current context $A\left(G_E^c\right)$, assuming the truth of the null hypothesis H_0. The pattern is accepted if H_0 is rejected. This happens if the *p-value* is under a critical significance value α which amounts to test if the observed intra-agreement $A\left(G_E^c\right)$ is outside the confidence interval $\mathrm{CI}^{1-\alpha}$ established using the distribution assumed under H_0.

H_0 corresponds to the baseline finding: the observed contextual intra-agreement is generated by the distribution of random subsets equally likely to occur, a.k.a. *Distribution of False Discoveries* (DFD, cf. [9]). We evaluate the *p-value* of the observed A against the distribution of random subsets of a cardinality equal to the size of the observed subgroup G_E^c. The subsets are issued by uniform sampling without replacement (since the observed subgroup encompasses distinct entities only) from the entity collection. Moreover, drawing samples only from the collection of subsets of size equal to $|G_E^c|$ allows to drive more judicious conclusions: the variability of the statistic A is impacted by the size of the considered subgroups, since smaller subgroups are more likely to observe low/high values of A. The same reasoning was followed in [33].

We define $\theta_k : F_k \to \mathbb{R}$ as the random variable corresponding to the observed intra-agreement A of k-sized subsets $S \in G_E$. I.e., for any $k \in [1, n]$ with $n = |G_E|$, we have $\theta_k(S) = A(S)$ and $F_k = \{S \in G_E \; s.t. \; |S| = k\}$. F_k is then the set of possible subsets which are equally likely to occur under the null hypothesis H_0. That is, $\mathbb{P}(S \in F_k) = \binom{n}{k}^{-1}$. We denote by $CI_k^{1-\alpha}$ the $(1-\alpha)$ confidence interval related to the probability distribution of θ_k under the null hypothesis H_0. To easily manipulate θ_k, we reformulate A using Equations (1)-(3):

$$A(S) = \frac{\sum_{e \in S} v_e}{\sum_{e \in S} w_e} \mid w_e = m_e \text{ and } v_e = m_e - \frac{1}{D_{\text{exp}}} \sum_{o_1, o_2 \in O^2} \delta_{o_1 o_2} \cdot \frac{m_e^{o_1} \cdot m_e^{o_2}}{(m_e - 1)} \quad (4)$$

Under the null hypothesis H_0 and the assumption that the underlying distribution of intra-agreements is a Normal distribution[5] $\mathcal{N}(\mu_k, \sigma_k^2)$, one can define

[4] This paradigm naturally raises the question of how to address the *multiple comparisons problem* [23]. This is a non-trivial task in our setting, and solving it requires an extension of the significant pattern mining paradigm as a whole: its scope is bigger than this paper. We provide a brief discussion in Appendix C.

[5] In the same line of reasoning of [7], one can assume that the underlying distribution can be derived from what prior beliefs the end-user may have on such distribution. If only the observed expectation μ and variance σ^2 are given as constraints which must hold for the underlying distribution, the maximum entropy distribution (*taking into account no other prior information than the given constraints*) is known to be the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ [5, p.413].

$CI_k^{1-\alpha}$ by computing $\mu_k = E[\theta_k]$ and $\sigma_k^2 = \text{Var}[\theta_k]$. Doing so requires either empirically calculating estimators of such moments by drawing a large number r of uniformly generated samples from F_k, or analytically deriving the formula of $E[\theta_k]$ and $\text{Var}[\theta_k]$. In the former case, the confidence interval $CI_k^{1-\alpha}$ endpoints are given by [17, p.9]: $\mu_k \pm t_{1-\frac{\alpha}{2},r-1}\sigma_k\sqrt{1+(1/r)}$, with μ_k and σ_k empirically estimated on the r samples, and $t_{1-\frac{\alpha}{2},r-1}$ the $(1-\frac{\alpha}{2})$ percentile of Student's t-distribution with $r-1$ degrees of freedom. In the latter case, (μ_k and σ_k are known/derived analytically), the $(1-\alpha)$ confidence interval can be computed in its most basic form, that is $CI_k^{1-\alpha} = [\mu_k - z_{(1-\frac{\alpha}{2})}\sigma_k, \mu_k + z_{(1-\frac{\alpha}{2})}\sigma_k]$ with $z_{(1-\frac{\alpha}{2})}$ the $(1-\frac{\alpha}{2})$ percentile of $\mathcal{N}(0,1)$.

However, due to the problem setting, empirically establishing the confidence interval is computationally expensive, since it must be calculated for each enumerated context. Even for relatively small behavioral datasets, this quickly becomes intractable. Alternatively, analytically deriving a computationally efficient form of $E[\theta_k]$ is notoriously difficult, given that $E[\theta_k] = \binom{n}{k}^{-1}\sum_{S\in F_k}\frac{\sum_{e\in S}v_e}{\sum_{e\in S}w_e}$ and $\text{Var}[\theta_k] = \binom{n}{k}^{-1}\sum_{S\in F_k}\left(\frac{\sum_{e\in S}v_e}{\sum_{e\in S}w_e} - E[\theta_k]\right)^2$.

Since θ_k can be seen as a weighted arithmetic mean, one can model the random variable θ_k as the ratio $\frac{V_k}{W_k}$, where V_k and W_k are two random variables $V_k : F_k \to \mathbb{R}$ and $W_k : F_k \to \mathbb{R}$ with $V_k(S) = \frac{1}{k}\sum_{e\in S}v_e$ and $W_k(S) = \frac{1}{k}\sum_{e\in S}w_e$. An elegant way to deal with a ratio of two random variables is to approximate its moments using the *Taylor series* following the line of reasoning of [12] and [26, p.351], since no easy analytic expression of $E[\theta_k]$ and $\text{Var}[\theta_k]$ can be derived.

Proposition 1 (An Approximate Confidence Interval $\widehat{CI}_k^{1-\alpha}$ for θ_k). *Given $k \in [1,n]$ and $\alpha \in]0,1]$ (significance critical value), $\widehat{CI}_k^{1-\alpha}$ is given by:*

$$\widehat{CI}_k^{1-\alpha} = \left[\widehat{E}[\theta_k] - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}, \widehat{E}[\theta_k] + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}\right] \tag{5}$$

with $\widehat{E}[\theta_k]$ a Taylor approximation for the expectation $E[\theta_k]$ expanded around (μ_{V_k}, μ_{W_k}), and $\widehat{\text{Var}}[\theta_k]$ a Taylor approximation for $\text{Var}[\theta_k]$ given by:

$$\widehat{E}[\theta_k] = \left(\frac{n}{k}-1\right)\frac{\mu_v}{\mu_w}\beta_w + \frac{\mu_v}{\mu_w} \qquad \widehat{\text{Var}}[\theta_k] = \left(\frac{n}{k}-1\right)\frac{\mu_v^2}{\mu_w^2}(\beta_v + \beta_w) \tag{6}$$

with:

$$\mu_v = \frac{1}{n}\sum_{e\in G_E} v_e \qquad \mu_w = \frac{1}{n}\sum_{e\in G_E} w_e \qquad n = |G_E|$$
$$\mu_{v^2} = \frac{1}{n}\sum_{e\in G_E} v_e^2 \qquad \mu_{w^2} = \frac{1}{n}\sum_{e\in G_E} w_e^2 \qquad \mu_{vw} = \frac{1}{n}\sum_{e\in G_E} v_e w_e$$

and:

$$\beta_v = \frac{1}{n-1}\left(\frac{\mu_{v^2}}{\mu_v^2} - \frac{\mu_{vw}}{\mu_v\mu_w}\right) \qquad \beta_w = \frac{1}{n-1}\left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v\mu_w}\right)$$

For a proof of these equations, see Appendix A.

Note that the complexity of the computation of the approximate confidence interval $\widehat{CI}_k^{1-\alpha}$ is $\mathcal{O}(n)$, with n the size of entities collection G_E.

4.1 Pruning the Search Space

Optimistic Estimate on Krippendorff's Alpha. To quickly prune unpromising areas of the search space, we define a tight optimistic estimate [18] on Krippendorff's alpha. Eppstein and Hirschberg [14] propose a smart *linear algorithm* `Random-SMWA`[6] to find subsets with maximum weighted average. Recall that A can be seen as a weighted average (cf. Equation (4)).

In a nutshell, `Random-SMWA` seeks to remove k values to find a subset of S having $|S| - k$ values with maximum weighted average. The authors model the problem as such: given $|S|$ values decreasing linearly with time, find the time at which the $|S| - k$ maximum values add to zero. In the scope of this work, given a user-defined support threshold σ_E on the minimum allowed size of context extents, k is fixed to $|S| - \sigma_E$. The obtained subset corresponds to the smallest allowed subset having support $\geq \sigma_E$ maximizing the weighted average quantity A. The `Random-SMWA` algorithm can be tweaked[7] to retrieve the smallest subset of size $\geq \sigma_E$ having analogously the minimum possible weighted average quantity A. We refer to the algorithm returning the maximum (resp. minimum) possible weighted average by $\texttt{RandomSMWA}^{\texttt{max}}$ (resp. $\texttt{RandomSMWA}^{\texttt{min}}$).

Proposition 2 (Upper and Lower Bounds for A). *Given $S \subseteq G_E$, minimum context support threshold σ_E, and the following functions:*

$$UB(S) = A\left(\texttt{RandomSMWA}^{\texttt{max}}(S, \sigma_E)\right) \qquad LB(S) = A\left(\texttt{RandomSMWA}^{\texttt{min}}(S, \sigma_E)\right)$$

we know that LB (resp. UB) is a lower (resp. upper) bound for A, i.e.:

$$\forall c, d \in \mathcal{D}_E \ : \ c \sqsubseteq d \ \wedge \ |G_E^c| \geq |G_E^d| \geq \sigma_E \Rightarrow LB(G_E^c) \leq A(G_E^d) \leq UB(G_E^c)$$

Using these results, we define the optimistic estimate for A as an interval bounded by the minimum and the maximum A measure that one can observe from the subsets of a given subset $S \subseteq G_E$, that is: $OE(S, \sigma_E) = [LB(S), UB(S)]$.

Nested Confidence Intervals for A. The desired property between two confidence intervals of the same significance level α related to respectively k_1, k_2 with $k_1 \leq k_2$ is that $CI_{k_1}^{1-\alpha}$ encompasses $CI_{k_2}^{1-\alpha}$. Colloquially speaking, larger samples lead to "narrower" confidence intervals. This property is intuitively plausible, since the dispersion of the observed intra-agreement for smaller samples is likely to be higher than the dispersion for larger samples. Having such a property allows to prune the search subspace related to a context c when traversing the search space downward if $OE(G_E^c, \sigma_E) \subseteq CI_{|G_E^c|}^{1-\alpha}$.

Proving $CI_{k_2}^{1-\alpha} \subseteq CI_{k_1}^{1-\alpha}$ for $k_1 \leq k_2$ for the exact confidence interval is nontrivial, since it requires to analytically derive $E[\theta_k]$ and $\text{Var}[\theta_k]$ for any $1 \leq k \leq n$. Note that the expected value $E[\theta_k]$ varies when k varies. We study such a property for the approximate confidence interval $\widehat{CI}_k^{1-\alpha}$.

[6] `Random-SMWA`: Randomized algorithm - Subset with Maximum Weighted Average.

[7] Finding the subset having the minimum weighted average is a dual problem to finding the subset having the maximum weighted average. To solve the former problem using `Random-SMWA`, we modify the values of v_i to $-v_i$ and keep the same weights w_i.

Proposition 3 (Minimum Cardinality Constraint for Nested Approximate Confidence Intervals). *Given a context support threshold σ_E and α.*

$$\textit{If } \sigma_E \geq C^{\alpha} = \frac{4n\beta_w^2}{z^2_{1-\frac{\alpha}{2}}(\beta_v + \beta_w) + 4\beta_w^2},$$
$$\textit{then } \forall k_1, k_2 \in \mathbb{N} : \sigma_E \leq k_1 \leq k_2 \Rightarrow \widehat{CI}^{1-\alpha}_{k_2} \subseteq \widehat{CI}^{1-\alpha}_{k_1}$$

Combining Propositions 1, 2 and 3, we formalize the pruning region property which answers: *when to prune the sub-search space under a context c?*

Corollary 1 (Pruning Regions). *Given a behavioral dataset $\mathcal{B}$, a context support threshold $\sigma_E \geq C^{\alpha}$, and a significance critical value $\alpha \in]0, 1]$. For any $c, d \in \mathcal{D}_E$ such that $c \sqsubseteq d$ with $|G^c_E| \geq |G^d_E| \geq \sigma_E$, we have:*

$$OE(G^c_E, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G^c_E|} \Rightarrow A(G^d_E) \in \widehat{CI}^{1-\alpha}_{|G^d_E|} \Rightarrow \textit{p-value}(d) > \alpha$$

Proofs. All proofs of propositions and properties can be found in Appendix A.

5 On Handling Variability of Outcomes Among Raters

In Section 4, we defined the confidence interval $CI^{1-\alpha}$ established over the DFD. By taking into consideration the variability induced by the selection of a subset of entities, such a confidence interval enables to avoid reporting subgroups indicating an intra-agreement likely (w.r.t. the critical value α) to be observed by a random subset of entities. For more statistically sound results, one should not only take into account the variability induced by the selection of subsets of entities, but also the variability induced by the outcomes of the selected group of individuals. This is well summarized by Hayes and Krippendorff [22]: "The obtained value of A is subject to random sampling variability—specifically variability attributable to the selection of units (i.e., entities) in the reliability data (i.e., behavioral data) and the variability of their judgments". To address these two questions, they recommend to employ a standard Efron & Tibshirani *bootstrapping approach* [13] to empirically generate the sampling distribution of A and produce an empirical confidence interval $\text{CI}^{1-\alpha}_{\text{bootstrap}}$.

Recall that we consider here a behavioral dataset $\mathcal{B}$ reduced to the outcomes of a selected group of individuals g. Following the bootstrapping scheme proposed by Krippendorff [22,28], the empirical confidence interval is computed by repeatedly performing the following steps: (1) resample n entities from G_E with replacement; (2) for each sampled entity, draw uniformly $m_e \cdot (m_e - 1)$ pairs of outcomes according to the distribution of the observed pairs of outcomes; (3) compute the observed disagreement and calculate Krippendorff's alpha on the resulting resample. This process, repeated b times, leads to a vector of bootstrap estimates (sorted in ascending order) $\hat{B} = [\hat{A}_1, \ldots, \hat{A}_b]$. Given the empirical distribution $\hat{B}$, the empirical confidence interval $\text{CI}^{1-\alpha}_{\text{bootstrap}}$ is defined by the percentiles of $\hat{B}$, i.e., $\text{CI}^{1-\alpha}_{\text{bootstrap}} = [\hat{A}_{\lfloor \frac{\alpha}{2} \cdot b \rfloor}, \hat{A}_{\lceil (1-\frac{\alpha}{2}) \cdot b \rceil}]$. We denote by $\text{MCI}^{1-\alpha}$ (Merged CI) the confidence interval that takes into consideration both $CI^{1-\alpha} = [\text{le}_1, \text{re}_1]$ and $\text{CI}^{1-\alpha}_{\text{bootstrap}} = [\text{le}_2, \text{re}_2]$. We have $\text{MCI}^{1-\alpha} = [\min(\text{le}_1, \text{le}_2), \max(\text{re}_1, \text{re}_2)]$.

6 A Branch-and-bound Solution: Algorithm DEvIANT

To detect exceptional contextual intra-group agreement patterns, we need to enumerate candidates $p = (g, c) \in (\mathcal{D}_I, \mathcal{D}_E)$. Both heuristic (e.g., beam search [31]) and exhaustive (e.g., GP-growth [32]) enumeration algorithms exist. We exhaustively enumerate all candidate subgroups while leveraging closure operators [15] (since A computation only depends on the extent of a pattern). This makes it possible to avoid redundancy and to substantially reduce the number of visited patterns. With this aim in mind, and since the data we deal with are of the same format as those handled in the previous work [3], we apply EnumCC to enumerate subgroups g (resp. c) in $\mathcal{D}_I$ (resp. $\mathcal{D}_E$). EnumCC follows the line of algorithm CloseByOne [29]. Given a collection G of records (G_E or G_I), EnumCC traverses the search space depth-first and enumerates only once all closed descriptions fulfilling the minimum support constraint σ. EnumCC follows a yield and wait paradigm (similar to Python's generators) which at each call yield the following candidate and wait for the next call. See Appendix B for details.

DEvIANT implements an efficient branch-and-bound algorithm to **D**iscover statistically significant **E**xceptional **I**ntra-group **A**greement pa**T**terns while leveraging closure, tight optimistic estimates and pruning properties. DEvIANT starts by selecting a group g of individuals. Next, the corresponding behavioral dataset $\mathcal{B}^g$ is established by reducing the original dataset $\mathcal{B}$ to elements concerning solely the individuals comprising G_I^g and entities having at least two outcomes. Subsequently, the bootstrap confidence interval $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$ is calculated.

Algorithm 1: DEvIANT$(\mathcal{B}, \sigma_E, \sigma_I, \alpha)$

Inputs : Behavioral dataset $\mathcal{B} = \langle G_I, G_E, O, o\rangle$, minimum support threshold σ_E of a context and σ_I of a group, and critical significance value α.
Output: Set of exceptional intra-group agreement patterns P.

$P \leftarrow \{\}$
foreach $(g, G_I^g, cont_g) \in \mathrm{EnumCC}(G_I, *, \sigma_I, 0, True)$ **do**
  $G_E(g) = \{e \in E \text{ s.t. } n_e^g \geq e\}$
  $\mathcal{B}^g = \langle G_E(g), G_I^g, O, o\rangle$
  $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}} = [\hat{A}_{\lfloor \frac{\alpha}{2}\cdot b\rfloor}, \hat{A}_{\lceil (1-\frac{\alpha}{2})\cdot b\rceil}]$ ▷ With $\hat{B} = [\hat{A}_1^g, ..., \hat{A}_b^g]$ computed on respectively b resamples of $\mathcal{B}^g$
  $\sigma_E^g = \max\left(C^\alpha\left(g\right), \sigma_E\right)$
  foreach $(c, G_E^c, cont_c) \in \mathrm{EnumCC}(G_E(g), *, \sigma_E^g, 0, True)$ **do**
    $\mathrm{MCI}^{1-\alpha}_{|G_E^c|} = \mathrm{merge}\left(\widehat{CI}^{1-\alpha}_{|G_E^c|}, \mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}\right)$
    if $\mathrm{OE}(G_E^c, \sigma_E^g) \subseteq \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$ **then**
      $\mathrm{cont}_c \leftarrow$ False ▷ Prune the unpromising search subspace under c
    else if $A^g(G_E^c) \notin \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$ **then**
      $p_{\mathrm{new}} \leftarrow (g, c)$
      if $\nexists p_{\mathrm{old}} \in P$ *s.t.* $\mathrm{ext}(p_{\mathrm{new}}) \subseteq \mathrm{ext}(p_{\mathrm{old}})$ **then**
        $P \leftarrow (P \cup p_{\mathrm{new}}) \setminus \{p_{\mathrm{old}} \in P \mid \mathrm{ext}(p_{\mathrm{old}}) \subseteq \mathrm{ext}(p_{\mathrm{new}})\}$
      $\mathrm{cont}_c \leftarrow$ False ▷ Prune the sub search space (generality concept)
return P

Table 3: Main characteristics of the behavioral datasets. $C^{0.05}$ represents the minimum context support threshold over which we have nested approximate CI property.

	$\lvert G_E \rvert$	$\mathcal{A}_E$ (Items-Scaling)	$\lvert G_I \rvert$	$\mathcal{A}_I$ (Items-Scaling)	Outcomes	Sparsity	$C^{0.05}$
EPD8[8]	4704	$1H + 1N + 1C$ (437)	848	$3C$ (82)	$3.1M$ (C)	78.6%	$\simeq 10^{-6}$
CHUS[9]	17350	$1H + 2N$ (307)	1373	$2C$ (261)	$3M$ (C)	31.2%	$\simeq 10^{-4}$
Movielens[10]	1681	$1H + 1N$ (161)	943	$3C$ (27)	$100K$ (O)	06.3%	$\simeq 0.065$
Yelp[11]	$127K$	$1H + 1C$ (851)	$1M$	$3C$ (6)	$4.15M$ (O)	0.003%	$\simeq 1.14$

Before searching for exceptional contexts, the minimum context support threshold σ_E is adjusted to $C^{\alpha}(g)$ (cf. Proposition 3) if it is lower than $C^{\alpha}(g)$. While in practice $C^{\alpha}(g) \ll \sigma_E$, we keep this correction for algorithm soundness. Next, contexts are enumerated by EnumCC. For each candidate context c, the optimistic estimate interval $OE(G_E^c)$ is computed (cf. Proposition 2). According to Corollary 1, if $OE(G_E^c, \sigma_E^g) \subseteq \text{MCI}^{1-\alpha}_{|G_E^c|}$, the search subspace under c can be pruned. Otherwise, $A^g(G_E^c)$ is computed and evaluated against $\text{MCI}^{1-\alpha}_{|G_E^c|}$. If $A^g(G_E^c) \notin \text{MCI}^{1-\alpha}_{|G_E^c|}$, then (g, c) is significant and kept in the result set P. To reduce the number of reported patterns, we keep only the most general patterns while ensuring that each significant pattern in $\mathcal{P}$ is represented by a pattern in P. This formally translates to: $\forall p' = (g', c') \in \mathcal{P} \setminus P : p\text{-}value^{g'}(c') \leq \alpha \Rightarrow \exists p = (g, c) \in P$ s.t. $\text{ext}(q) \subseteq \text{ext}(p)$, with $\text{ext}\,(q = (g', c')) \subseteq \text{ext}\,(p = (g, c))$ defined by $G_I^{g'} \subseteq G_I^g$ and $G_E^{c'} \subseteq G_E^c$. This is based on the following postulate: the end-user is more interested by exceptional (dis-)agreement within larger groups and/or for larger contexts rather than local exceptional (dis-)agreement. Moreover, the end-user can always refine their analysis to obtain more fine-grained results by re-launching the algorithm starting from a specific context or group.

7 Empirical Evaluation

Our experiments aim to answer the following questions: **($\mathbf{Q}_1$)** How well does the Taylor-approximated CI approach the empirical CI? **($\mathbf{Q}_2$)** How efficient is the Taylor-approximated CI and the pruning properties? **($\mathbf{Q}_3$)** Does DEvIANT provide interpretable patterns? Source code and data are available on our companion page: https://github.com/Adnene93/Deviant.

Datasets. Experiments were carried on four real-world behavioral datasets (cf. Table 3): two voting (EPD8 and CHUS) and two rating datasets (Movielens and Yelp). Each dataset features entities and individuals described by attributes that are either categorical (C), numerical (N), or categorical augmented with a taxonomy (H). We also report the equivalent number of items (in an itemset language) corresponding to the descriptive attributes (ordinal scaling [16]).

[8] Eighth European Parliament Voting Dataset (04/10/18).

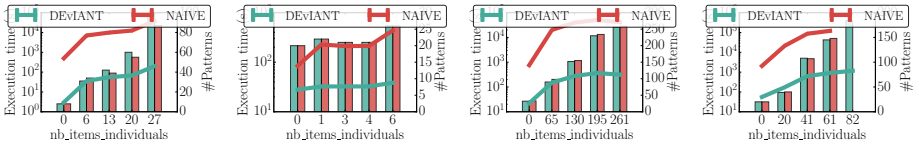
[9] 102^{nd}-115^{th} congresses of the US House of Representatives (Period: 1991-2015).

[10] Movie review dataset - https://grouplens.org/datasets/movielens/100k/.

[11] Social network dataset - https://www.yelp.com/dataset/challenge (25/04/17).

Table 4: Coverage error between empirical CIs and Taylor CIs.

$\mathcal{B}$	μ_{err}	σ_{err}	$\|\|\mathcal{B}$	μ_{err}	σ_{err}	$\|\|\mathcal{B}$	μ_{err}	σ_{err}	$\|\|\mathcal{B}$	μ_{err}	σ_{err}
CHUS	0.007	0.004	$\|\|$EPD8	0.007	0.004	$\|\|$Movielens	0.0075	0.0045	$\|\|$Yelp	0.007	0.004



(a) Movielens **(b)** Yelp **(c)** EPD8 **(d)** CHUS

Fig. 2: Comparison between DEvIANT and `Naive` when varying the size of the description space $\mathcal{D}_I$. Lines correspond to the execution time and bars correspond to the number of output patterns. Parameters: $\sigma_E = \sigma_I = 1\%$ and $\alpha = 0.05$.

Q_1. First, we evaluate to what extent the empirically computed confidence interval approximates the confidence interval computed by Taylor approximations. We run 1000 experiments for subset sizes k uniformly randomly distributed in $[1, n = |G_E|]$. For each k, we compute the corresponding Taylor approximation $\widehat{CI}_k^{1-\alpha} = [a^T, b^T]$ and empirical confidence interval $\text{ECI}_k^{1-\alpha} = [a^E, b^E]$. The latter is calculated over 10^4 samples of size k from G_E, on which we compute the observed A which are then used to estimate the moments of the empirical distribution required for establishing $\text{ECI}_k^{1-\alpha}$. Once both CIs are computed, we measure their distance by Jaccard index. Table 4 reports the average μ_{err} and the standard deviation σ_{err} of the observed distances (coverage error) over the 1000 experiments. Note that the difference between the analytic Taylor approximation and the empirical approximation is negligible ($\mu_{err} < 10^{-2}$). Therefore, the CIs approximated by the two methods are so close, that it does not matter which method is used. Hence, the choice is guided by the computational efficiency.

Q_2. To evaluate the pruning properties' efficiency (**(i)** Taylor-approximated CI, **(ii)** optimistic estimates and **(iii)** nested approximated CIs), we compare DEvIANT with a `Naive` approach where the three aforementioned properties are disabled. For a fair comparison, `Naive` pushes monotonic constraints (minimum support threshold) and employs closure operators while empirically estimating the CI by successive random trials from F_k. In both algorithms we disable the bootstrap $\text{CI}_{\text{bootstrap}}^{1-\alpha}$ computation, since its overhead is equal for both algorithms. We vary the description space size related to groups of individuals $\mathcal{D}_I$ while considering the full entity description space. Figure 2 displays the results: DEvIANT outperforms `Naive` in terms of runtime by nearly two orders of magnitude while outputting the same number of the desired patterns.

Figure 3 reports the performance of DEvIANT in terms of runtime and number of output patterns. When varying the description space size, DEvIANT requires more time to finish. Note that the size of individuals search space $\mathcal{D}_I$ substantially affects the runtime of DEvIANT. This is mainly because larger

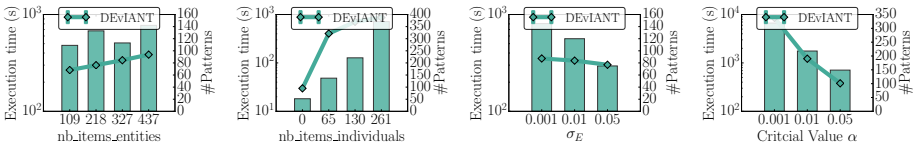


Fig. 3: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$.

Table 5: All the exceptional consensual/conflictual subjects among **Republican Party** representatives (selected upfront, i.e. G_I restricted over members of Republican party) in the 115^{th} congress of the US House of Representatives. $\alpha = 0.01$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	p-value	IA
p_1	Republicans	20.11 Government and Administration issues	0.83	0.32	$<.001$	Conflict
p_2	Republicans	5 Labor	0.83	0.63	$<.01$	Conflict
p_3	Republicans	20.05 Nominations and Appointments	0.83	0.92	$<.001$	Consensus

$\mathcal{D}_I$ leads to more candidate groups of individuals g which require DEvIANT to: (i) generate $\text{CI}^{1-\alpha}_{\text{bootstrap}}$ and (ii) mine for exceptional contexts c concerning the candidate group g. Finally, when α decreases, the execution time required for DEvIANT to finish increases while returning more patterns. This may seem counter-intuitive, since fewer patterns are significant when α decreases. It is a consequence of DEvIANT considering only the most general patterns. Hence, when α decreases, DEvIANT goes deeper in the context search space: much more candidate patterns are tested, enlarging the result set. The same conclusions are found on the Yelp, Movielens, and CHUS datasets (cf. Appendix D).

Q_3. Table 5 reports exceptional contexts observed among House Republicans during the 115^{th} Congress. Pattern p_1, illustrated in Figure 4, highlights a collection of voting sessions addressing Government and Administrative issues where a clear polarization is observed between two clusters of Republicans. A roll call vote in this context featuring significant disagreement between Republicans is "**House Vote 417**" (cf. https://projects.propublica.org/represent/votes/115/house/1/417) which was closely watched by the media (Washington Post: https://wapo.st/2W32I9c; Reuters: https://reut.rs/2TF0dgV).

Table 6 depicts patterns returned by DEvIANT on the Movielens dataset. Pattern p_2 reports that "Middle-aged Men" observe an intra-group agreement significantly higher than overall, for movies labeled with both adventure and musical genres (e.g., The Wizard of Oz (1939)).

8 Conclusion and Future Directions

We introduce the task to discover statistically significant exceptional contextual intra-group agreement patterns. To efficiently search for such patterns, we devise DEvIANT, a branch-and-bound algorithm leveraging closure operators,

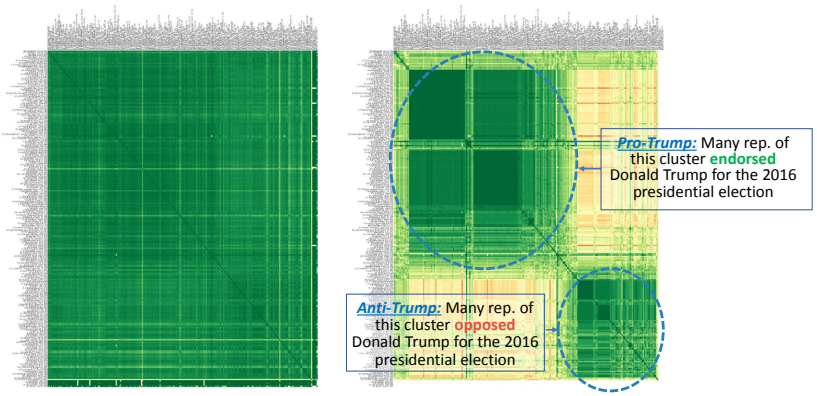


Fig. 4: Similarity matrix between Republicans, illustrating Pattern p_1 from Table 5. Each cell represents the ratio of voting sessions in which Republicans agreed. Green cells report strong agreement; red cells highlight strong disagreement.

Table 6: Top-3 exceptionally consensual/conflictual genres between Movielens raters, α=0.01. Patterns are ranked by absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	p-value	IA
p_1	Old	1.Action & 2.Adventure & 6.Crime Movies	-0.06	-0.29	< 0.01	Conflict
p_2	Middle-aged Men	2.Adventure & 12.Musical Movies	0.05	0.21	< 0.01	Consensus
p_3	Old	4.Children & 12.Musical Movies	-0.06	-0.21	< 0.01	Conflict

approximate confidence intervals, tight optimistic estimates on Krippendorff's Alpha measure, and the property of nested CIs. Experiments demonstrate DEvIANT's performance on behavioral datasets in domains ranging from political analysis to rating data analysis. In future work, we plan to (i) investigate how to tackle the multiple comparison problem [21], (ii) investigate intra-group agreement which is exceptional w.r.t. all individuals *over the same context*, and (iii) integrate the option to choose which kind of exceptional consensus the end-user wants: is the exceptional consensus caused by common preference or hatred for the context-related entities? All this is to be done within a comprehensive framework and tool (prototype available at `http://contentcheck.liris.cnrs.fr`) for behavioral data analysis alongside exceptional inter-group agreement pattern discovery implemented in [3].

Acknowledgments. This work has been partially supported by the project *ContentCheck* **ANR-15-CE23-0025** funded by the French National Research Agency. The authors would like to thank the reviewers for their valuable remarks. They also warmly thank Arno Knobbe, Simon van der Zon, Aimene Belfodil and Gabriela Ciuperca for interesting discussions.

A Appendix: Proofs

Recall that $\theta_k : F_k \to \mathbb{R}$ is the random variable corresponding to the observed intra-agreement A (Krippendorff's alpha) of subsets $S \in G_E$ of size k. I.e., for any $k \in [1, n]$ with $n = |G_E|$ we have $\theta_k(S \in F_k) = A(S)$ and $F_k = \{S \in G_E \; s.t. \; |S| = k\}$. Then, F_k is the set of possible outcomes which are equally likely to occur under the null hypothesis H_0. We let n denote the number of records in G_E (i.e., $|G_E| = n$). Each record $e \in G_E$ is associated with a value v_e and w_e. The quantity θ_k can be expressed as a ratio $\frac{V_k}{W_k}$, where V_k, W_k are two random variables $V_k : F_k \to \mathbb{R}$ and $W_k : F_k \to \mathbb{R}$ with $V_k(S) = \frac{1}{k}\sum_{e\in S} v_e$ and $W_k(S) = \frac{1}{k}\sum_{e\in S} w_e$.

Proof (Proposition 1). For any $f(x, y)$, the bivariate second order Taylor expansion about any $\lambda = (\lambda_x; \lambda_y)$ is:[15]

$$\begin{aligned} f(x,y) &= f(\lambda) + f'_x(\lambda)(x-\lambda_x) + f'_y(\lambda)(y-\lambda_y) \\ &+ \frac{1}{2}\left(f''_{xx}(\lambda)(x-\lambda_x)^2 + 2f''_{xy}(\lambda)(x-\lambda_x)(y-\lambda_y) + f''_{yy}(\lambda)(y-\lambda_y)^2\right) + \epsilon \end{aligned} \tag{7}$$

where ϵ is a remainder of smaller order than the term of the equation.
An approximation of the expectation $E[f(x, y)]$ expanded around $\lambda = (\lambda_x; \lambda_y)$ is:

$$E[f(x,y)] \approx f(\lambda) + \frac{1}{2}\left[f''_{xx}(\lambda)\mathrm{Var}[X] + 2f''_{xy}(\lambda)\mathrm{Cov}[X,Y] + f''_{yy}(\lambda)\mathrm{Var}[Y]\right]$$

Given that $f(x, y) = \frac{x}{y}$ and using the fact that $E[X - \mu_x] = 0$ (which is valid for both V and W), we have: $\mathrm{Var}[X] = E[(X - \mu_x)^2]$ and $\mathrm{Cov}[X, Y] = (X - \mu_x)(Y - \mu_y)$. We can derive an approximation of $E[\theta_k] = E[\frac{V_k}{W_k}]$ around (μ_{V_k}, μ_{W_k}):

$$E[\theta_k] = E[\frac{V_k}{W_k}] = E[f(V_k, W_k)] \approx \frac{\mu_{V_k}}{\mu_{W_k}} - \frac{\mathrm{Cov}[V_k, W_k]}{\mu^2_{W_k}} + \frac{\mathrm{Var}[W_k]\mu_{V_k}}{\mu^3_{V_k}} \tag{8}$$

The formulas of $E[V_k]$ (resp. $E[W_k]$) and $\mathrm{Var}[V_k]$ (resp. $V[W_k]$) can be derived analytically. We denote by μ_v (resp. μ_w) the arithmetic mean of the values (resp. weights) corresponding to each entity $e \in G_E$, i.e.: $\mu_v = \frac{1}{n}\sum_{e\in G_E} v_e$ and $\mu_w = \frac{1}{n}\sum_{e\in G_E} w_e$ with $n = |G_E|$.

$$E[V_k] = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \frac{1}{k}\sum_{e\in S} v_e = \frac{1}{n}\sum_{e\in G_E} v_e = \mu_v \tag{9}$$

$$\begin{aligned} \mathrm{Var}[V_k] &= \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - E[V_k]\right)^2 = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - \mu_v\right)^2 \\ &= \frac{1}{k}\left(\frac{n}{n-1}\left(\mu_{v^2} - \mu_v^2\right)\right) - \frac{1}{n-1}\left(\mu_{v^2} - \mu_v^2\right) \text{ with } \mu_{v^2} = \frac{1}{n}\sum_{e\in G_E} v_e^2 \end{aligned} \tag{10}$$

[15] a concise lecture note follows the same reasoning and explains the derivations; see http://www.stat.cmu.edu/~hseltman/files/ratio.pdf

The same reasoning applies to compute the expected value and the variance related to W_k:

$$E[W_k] = \frac{1}{n} \sum_{e \in G_E} w_e = \mu_w \tag{11}$$

$$\begin{aligned} \mathrm{Var}[W_k] &= \frac{1}{\binom{n}{k}} \sum_{S \in F_k} \left(\frac{1}{k} \sum_{e \in S} w_e - E[W_k] \right)^2 \\ &= \frac{1}{k} \left(\frac{n}{n-1} \left(\mu_{w^2} - \mu_w^2 \right) \right) - \frac{1}{n-1} \left(\mu_{w^2} - \mu_w^2 \right) \text{ with } \mu_{w^2} = \frac{1}{n} \sum_{e \in G_E} w_e^2 \end{aligned} \tag{12}$$

We now derive the formula for $\mathrm{Cov}(V_k, W_k)$. The same line of reasoning for the computation of the variance of V_k and W_k applies. We obtain:

$$\begin{aligned} \mathrm{Cov}[V_k, W_k] &= \frac{1}{\binom{n}{k}} \sum_{S \in F_k} \left(\frac{1}{k} \sum_{e \in S} v_e - E[V_k] \right) \left(\frac{1}{k} \sum_{e \in S} w_e - E[W_k] \right) \\ &= \frac{1}{k} \left(\frac{n}{n-1} \left(\mu_{vw} - \mu_v \mu_w \right) \right) - \frac{1}{n-1} \left(\mu_{vw} - \mu_v \mu_w \right) \\ &\text{with } \mu_{vw} = \frac{1}{n} \sum_{e \in G_E} w_e v_e \end{aligned} \tag{13}$$

Using Equations *(9)*, *(10)*, *(11)*, *(12)*, *(13)*, we derive the approximation of $E[\theta_k]$ after simplifications of *(8)*:

$$E[\theta_k] \approx \widehat{E}[\theta_k] = \left(\frac{n}{k} - 1 \right) \frac{\mu_v}{\mu_w} \beta_w + \frac{\mu_v}{\mu_w} \text{ with } \beta_w = \frac{1}{n-1} \left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \tag{14}$$

The same reasoning applies to approximate $\mathrm{Var}[\theta_k]$ using Taylor expansions. We will confine ourselves to a first-order Taylor expansion around (μ_v, μ_w) to make the analytic derivation of the approximation of $\mathrm{Var}[\theta_k]$ feasible. The same observation has been made by [25,12] and [26, p. 351] to approximate the variance for a ratio random variable. We obtain:

$$\mathrm{Var}[\theta_k] = \mathrm{Var}[f(V_k, W_k)] \approx \frac{\mathrm{Var}[V_k]}{\mu_{W_k}^2} - 2 \frac{\mu_{V_k} \mathrm{Cov}[V_k, W_k]}{\mu_{W_k}^3} + \frac{\mu_{V_k}^2 \mathrm{Var}[W_k]}{\mu_{W_k}^4} \tag{15}$$

After simplifications and by using the same line of reasoning when deriving the expected value approximation reported in Equation (14), we obtain:

$$\begin{aligned} \mathrm{Var}[\theta_k] &\approx \widehat{\mathrm{Var}}[\theta_k] = \left(\frac{n}{k} - 1 \right) \frac{\mu_v^2}{\mu_w^2} \left(\beta_v + \beta_w \right) \\ &\text{with } \beta_w = \frac{1}{n-1} \left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \text{ and } \beta_v = \frac{1}{n-1} \left(\frac{\mu_{v^2}}{\mu_v^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \end{aligned} \tag{16}$$

We denote by $\widehat{CI}_k^{1-\alpha}$ the approximate confidence interval calculated using the approximations from Equations (14) and (16) of the expected value $\widehat{E}[\theta_k]$ and the variance $\widehat{\text{Var}}[\theta_k]$, respectively. This results in:

$$\widehat{CI}_k^{1-\alpha} = \left[\widehat{E}[\theta_k] - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}, \widehat{E}[\theta_k] + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}\right]$$

It is worth mentioning that the complexity of the computation of this approximate confidence interval is linear to the size n. □

Proof (Proposition 2). To simplify the text, we will omit σ_E as a parameter in the proof and keep in mind that we consider the minimum support threshold σ_E. Given that $c \sqsubseteq d$, with c, d two descriptions from $\mathcal{D}$, we have $G_E^d \subseteq G_E^c$. The proposition stems from the fact that:

1. $A(G_E^c) \leq UB(G_E^c)$, since `RandomSMWA`$^{\texttt{max}}$ computes the subset S_{max}^c having the maximum weighted average A as proven by Epstein and Hirschberg [14].
2. UB is monotonic w.r.t. the partial order $\subseteq$ between sets. That is:

$$\forall S, S' \subseteq G_E \ : \ S' \subseteq S \Rightarrow UB(S') \leq UB(S)$$

This can be proven by reductio ad absurdum. We denote by $S'_{\text{max}} \subseteq S'$ (resp. $S_{\text{max}} \subseteq S$) the optimal subset of S' (resp. S) having its size $\geq \sigma_E$ and the maximum possible weighted average A. Suppose that $\exists S, S' \subseteq G_E \ : \ S' \subseteq S \wedge UB(S') > UB(S)$ $(A(S'_{\text{max}}) > A(S_{\text{max}}))$. Since $S' \subseteq S$, this means that there is another subset in S, namely S'_{max}, that observes a greater weighted average A than the actual optimal subset S_{max}, which is absurd.

From properties 1. and 2. we have: $A(G_E^d) \leq UB(G_E^d) \leq UB(G_E^c)$. The same reasoning holds to prove that LB is a lower bound. □

Proof (Proposition 3). In order to prove the desired property for the approximate confidence intervals, we first must determine if the variance decreases when k increases.

$$k_1, k_2 \in \mathbb{N} : \text{ if } k_1 \leq k_2 \Rightarrow \widehat{\text{Var}}[\theta_{k_1}] \geq \widehat{\text{Var}}[\theta_{k_2}] \tag{17}$$

From Equation (16), $\widehat{\text{Var}}[\theta_k] = \left(\frac{n}{k} - 1\right)\frac{\mu_v^2}{\mu_w^2}(\beta_v + \beta_w)$. Given that $\frac{n}{k} - 1$ is a decreasing function w.r.t. k, proving Equation (17) requires that $\beta_v + \beta_w$ is a positive quantity. This stems from the fact that the original formula of the approximate variance given in Equation (15) is positive. This can be proved by a direct application of the Covariance inequality [36, p. 149], which itself is an application of the Cauchy-Schwarz inequality [38]. Since $\beta_v + \beta_w$ is of the same sign of Equation (16), we have $\beta_v + \beta_w \geq 0$. For the sake of a self-contained proof. We give the proof of this assertion below:

From Equations (15) and (16), we have: $\beta_v + \beta_w$ is of the same sign of:

$$\frac{\text{Var}[V_k]}{\mu_{V_k}^2} - 2\frac{\text{Cov}[V_k, W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\text{Var}[W_k]}{\mu_{W_k}^2} \tag{18}$$

From the Covariance inequality, we have $\text{Cov}[V_k, W_k] \leq \sigma[V_k]\sigma[W_k]$ with $\sigma^2[V_k] = \text{Var}[V_k]$ and $\sigma^2[W_k] = \text{Var}[W_k]$, hence Equation (18) is greater than:

$$\begin{aligned}
&\frac{\sigma^2[V_k]}{\mu^2_{V_k}} - 2\frac{\sigma[V_k]\sigma[W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\sigma^2[W_k]}{\mu^2_{W_k}} \\
&= \frac{\sigma[V_k]}{\mu_{V_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) - \frac{\sigma[W_k]}{\mu_{W_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) \\
&= \left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right)^2 \\
&\geq 0
\end{aligned}$$

Hence $\beta_v + \beta_w \geq 0$, which confirms that the variance is decreasing under increasing size k, as stated in Equation (17).

Recall that, by approximation, we want to ensure that for $\sigma_E \leq k_1 \leq k_2$ with σ_E a threshold on the context support, we have $\widehat{CI}^{1-\alpha}_{k_2} \subseteq \widehat{CI}^{1-\alpha}_{k_1}$. Hence, we need to find the minimum σ_E above which such property is valid. This amounts to finding a lower bound for σ_E such that:

$$z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_{k_1}]} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_{k_2}]} \geq \left|\widehat{E}[\theta_{k_1}] - \widehat{E}[\theta_{k_2}]\right| \tag{19}$$

Using the definitions of $\widehat{\text{Var}}[\theta_k]$ and $\widehat{E}[\theta_k]$ from Equations (14) and (16), the Equation (19) can be rewritten to:

$$\left(\sqrt{\frac{n}{k_1} - 1} + \sqrt{\frac{n}{k_2} - 1}\right) \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta^2_w}}$$

Since $\sigma_E \leq k_1 \leq k_2$, we require that:

$$2\sqrt{\frac{n}{\sigma_E} - 1} \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta^2_w}}$$

After simplifications, we obtain that σ_E must satisfy the following constraint:

$$\sigma_E \geq C^\alpha = \frac{4n\beta^2_w}{z^2_{1-\frac{\alpha}{2}}(\beta_v + \beta_w) + 4\beta^2_w} \qquad \square$$

Proof (Corollary 1). The proof is straightforward. From Proposition 2, we have that for any $c, d \in \mathcal{D}_E$ s.t. $c \sqsubseteq d$, if $G^c \geq G^d \geq \sigma_E$ then:

$$A\left(G^d_E\right) \in OE\left(G^c_E, \sigma_E\right) \tag{20}$$

From Proposition 3, if $\sigma_E \geq C^\alpha$ we have:

$$CI^{1-\alpha}_{|G^c_E|} \subseteq \widehat{CI}^{1-\alpha}_{|G^d_E|} \tag{21}$$

From Equations (20) and (21) and the fact that $OE(G^c_E, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G^c_E|}$, it follows that $A\left(G^d_E\right) \in OE\left(G^c_E, \sigma_E\right) \subseteq \widehat{CI}^{1-\alpha}_{|G^c_E|} \subseteq \widehat{CI}^{1-\alpha}_{|G^d_E|}$, hence *p-value(d)* $> \alpha$. $\square$

B Appendix: Enumeration Algorithm

Given a collection of records G whose descriptive attributes are $\mathcal{A} = \{a_1, \ldots, a_l\}$ which can be Boolean, numerical, or categorical, potentially organized among a taxonomy. Attributes $\mathcal{A}$ allow to structure the search space $\mathcal{D}$ by considering descriptions $d \in \mathcal{D}$, which are conjunctions of conditions over the attributes' domains of interpretation. A condition over a categorical attribute is an equality test while a condition over a numerical attributes is a membership test in an interval. By G^d we denote the set of records of G covered by the description d.

The EnumCC algorithm enumerates once and only once all closed descriptions whose associated subgroups fulfill the minimum support constraint σ. The algorithm follows the same reasoning of most common SD algorithms and goes in the same line of the CloseByOne Algorithm (CbO) [30] and the Divide-And-Conquer Algorithm [4]. It traverses the search lattice $\mathcal{D}$ in a top-down, DFS fashion starting from the most general description $*$ whose support is the entire collection G. It proceeds by atomic refinements to progress, step by step, toward more specific descriptions. This is enabled by a refinement operator denoted η_j for the j^{th} attribute. η_j keeps all conditions related to attributes a_i for $i \neq j$ intact, and refines only the j^{th} condition. If the condition is related to a numerical attribute, a minimal change to the left or right is performed [24]. If the condition is related to a categorical attribute, return an equality test for all possible values of the domain (if the condition was never refined before), otherwise no refinement is possible. If the attribute is an HMT (categorical attribute augmented with a taxonomy) only one tag is refined to its child or an additional tag is appended [3]. In a nutshell, for each parameter description d, EnumCC starts by assessing if the subgroup G^d is valid ($|G^d| \geq \sigma$). In this case, the closed description `closure_d` is computed and returned only if the canonicity test is passed (cf. [16, p.66-68]). The description `closure_d` corresponds to the tightest description of G^d (maximal in terms of the partial order $\sqsubseteq$ on descriptions in

Algorithm 2: EnumCC(G, d, σ_G, f, cnt)

Inputs : G is the collection of records, each encompassing m attributes, d is a description from $\mathcal{D}$, σ_G is a support threshold, $f \in [1, m]$ is a refinement flag, `cnt` is a Boolean.

Output: yields all closed descriptions, i.e. $clo[\mathcal{D}] = \{clo(d) \text{ s.t. } d \in \mathcal{D}\}$

```
1  if |G^d| ≥ σ then
2      closure_d ← δ(G^d)                ▷ compute the most specific description of G^d
3      if d ⋖_f closure_d then
4          cnt_c ← copy(cnt)             ▷ cnt_c value can be modified by a caller algorithm
5          yield (closure_d, G^closure_d, cnt_c)   ▷ yield results and wait for next call
6          if cnt_c then
7              foreach j ∈ [f, l] do
8                  foreach d' ∈ η_j(closure_d) do
9                      foreach (nc, G^nc, cnt_nc) ∈ EnumCC(G, d', σ_G, j, cnt_c) do
10                         yield (nc, G^nc, cnt_nc)
```

$\mathcal{D}$) which is the conjunction of all descriptions (conjunction of conditions) related to the records $g \in G^d$. Next, if the caller-algorithm allows the algorithm to continue (Boolean `cnt_c` kept `True`), the description `closure_d` is refined by starting from the last refined attribute (pointed out by the flag $f \in [1..l]$), since refining preceding attributes will certainly cause the next canonicity test to fail causing the algorithm to backtrack. Eventually, a recursive call is done to explore the sub-search space related to d (`closure_d`).

C Appendix: Multiple Comparisons Problem

In what follows, each pattern $H_i = (g_i, c_i)$ is seen as a hypothesis test which returns a p-value p_i. Recall that, in this paper, the list of hypotheses to test corresponds to the full search space $L = \{(g, c) \in \mathcal{D}_I \times \mathcal{D}_E : |G_I^g| \geq \sigma_I \text{ and } |G_E^c| \geq \sigma_E\}$ where g (resp. c) is a closed description (i.e. the maximum description w.r.t. $\sqsubseteq$) in the equivalence class $[g]$ (resp. $[c]$) of descriptions having their extent equal to G_I^g (resp. G_E^c), i.e. $[g] = \{g' \in \mathcal{D}_I \text{ s.t. } G_I^{g'} = G_I^g\}$ (resp. $[c] = \{c' \in \mathcal{D}_E \text{ s.t. } G_E^{c'} = G_E^c\}$). Having this in mind, in what follows, the content of L is shortly denoted by $L = \{H_1, \ldots, H_\omega\}$ and comprises ω hypotheses. Hypotheses in L are ordered by their p-values $\{p_1, \ldots, p_\omega\}$ where $p_i = p\text{-}value^{g_i}(c_i)$.

The Multiple Comparisons Problem (MCP) [23] is a critical issue in significant pattern mining [21]. In a nutshell, given the critical value α which roughly corresponds to the probability of type 1 error (rejecting a true null hypothesis which is equivalent to accepting a spurious pattern), it is to be expected that $\omega \cdot \alpha$ hypotheses will erroneously pass the test, i.e., $\omega \cdot \alpha$ hypotheses suffer a type 1 error. The classic approach to deal with the MCP is to control the *family wise error rate* (FWER), which is the probability of accepting at least one false discovery. Other approaches control the *false discovery rate* (FDR), which corresponds to the expected proportion of false discoveries. We give an overview of relevant existing approaches that deal with the MCP and point out why using them in our setting is a non-trivial task. For a survey on methods dealing with the MCP, we refer the interested reader to [21].

The most famous method to control FWER at $\leq \gamma$ (typically 0.05) is Bonferroni adjustments [11]. The critical α used to test the significance of a pattern is adjusted to $\frac{\gamma}{\omega}$ so as to have FWER at $\leq \gamma$ with ω the number of all patterns to test in L. The problem with this approach is that when ω is huge[16], Bonferroni adjusts α to a value very close to 0. This leads to a high number of false negatives as most interesting pattern will be considered spurious (high Type 2 error rate). Clearly, ω is unknown and needs, in the most trivial way, to be bounded by a quantity ω_0 which is **larger** than ω. Usually, ω_0 corresponds to the maximum size of the search space: it is equal to $2^{\#\text{items}}$ in the case of an itemset dataset. Webb gives a bound [44] on the size of the search space when dealing with the MCP in attribute-value datasets when the description length is bounded. Using this reasoning without bounding the description length and considering the specification of each attribute (numerical, categorical, ...), in the smallest of our datasets (Movielens; see Table 3) we have $\omega_0 = 72\,349\,200$. This requires α to be equal to 6.92×10^{-10} for the FWER to be at ≤ 0.05. All the other datasets require α to be $\leq 10^{-76}$ when bounding ω with the size of the search space. Clearly, such settings for α prohibit the discovery of any meaningful information from the datasets, which cannot possibly be the desired effect of attempts at solving the MCP.

[16] Which is the case in the general setting of pattern mining even if we consider only closed patterns satisfying the support size threshold constraint.

Several techniques exist in the literature to relax the requirements on α while ensuring a FWER at $\leq \gamma$ in order to increase the statistical power:

1. Terada et al. [41,40] propose the LAMP technique, relying on Tarone's Exclusion Principle (TEP) [39]. This principle stipulates that in the list of m hypotheses in L to be tested, one must ignore *untestable patterns* for multiple comparisons. A pattern H_i is said to be *untestable* if the **lower bound of its p-value**, denoted p_i^*, is under the adjusted $\alpha = \frac{\gamma}{m}$. Terada et al. [41] proposed this lower bound p_i^* for the particular task of finding significant rules[17] [43] where significance is commonly assessed using a Fisher exact test [19,20], since a 2×2 contingency table is available. The lower bound p_i^* computation depends on this contingency table. Clearly, there is no trivial mapping of our problem to the problem of finding significant rules. Hence, adapting the LAMP algorithm to have an efficient branch and bound technique, incorporating both the proposed bounds in this work (the DEvIANT algorithm) and LAMP reasoning, is clearly a daunting task that requires an in-depth investigation and a new devoted approach which is beyond the scope of this work.
2. Similarly, most of the existing work measuring the interestingness of patterns with statistical significance while efficiently handling the MCP, deals with the significant rule discovery setting [42,27,34,37]. Some of these methods [42,34,37] rely on the Westfall-Young permutation testing method [46] to increase statistical power. Still, no straightforward application of these techniques in our setting is possible: these methods perform random permutations on the class label, and no class label is given in the problem addressed in our work.
3. Other state-of-the-art techniques follow a multi-stage procedure [21] to tackle the MCP. A first step constrains L to a subset of patterns (e.g., testable under TEP). A subsequent post-processing phase controls the FWER [44] or FDR [44,27]. For example, Webb [44] proposes to divide the data into Exploratory and Holdout data. Hypotheses are sought by analyzing solely the exploratory data. Eventually, a constrained number of patterns are found which are validated against the holdout data. In our setting, one needs to investigate how to divide the data into these two parts, since we have two dimensions: context space and group space. In this configuration, a question of crucial importance must be answered: do we need to consider each group independently and divide the entities dataset (defining the context space) into exploratory vs holdout data for each group? Or do we need to jointly consider both these dimensions? This clearly requires a thorough investigation to avoid proposing a naive solution.
4. Layered critical values [45,2] propose to consider a varying adjustment factor for each level of the search space as long as the sum of all critical values is not above γ. This requires:

[17] Each record in the underlying dataset is associated with a binary target label and the objective is to find rules that have significant association with one of the two labels.

- estimating the size of each level (which could be done by following the reasoning of Webb in [45]);
- identifying what is a level of the search space: do we consider levels jointly between group and context search space?

Choosing joint consideration in the latter bullet point implies ignoring (most of the time) the level-1 groups in the search space: the level will grow in size after considering all the contexts corresponding to the group characterizing the whole collection of individuals. Otherwise, the question raised in the former bullet point needs to be answered to provide an appropriate algorithm. Furthermore, combining the layered critical values along with DEvIANT is not straightforward as it requires re-investigation of the proposed pruning properties.

As we can see, several fundamental questions remain to be answered before one could incorporate a solution to the MCP in the task of finding significant exceptional contextual intra-group agreement patterns. We argue that the scope of this problem is bigger than the ECMLPKDD 2019 publication at hand; it is a non-trivial task that deserves proper attention in the wider context of the significant pattern mining paradigm. We plan to investigate this in future work, and expect that the scope is too wide to fit within a single conference paper; a proper exploration probably requires a journal-length publication.

D Appendix: Additional Experiments

D.1 Performance evaluation

Additional experiments reporting the execution time and the number of reported significant patterns by DEvIANT on Movielens, Yelp, CHUS, and EPD8. In these experiments we also study the overhead induced by the computation of the bootstrapping confidence interval required to handle the variability of outcomes and evaluated for each generated group of individuals.

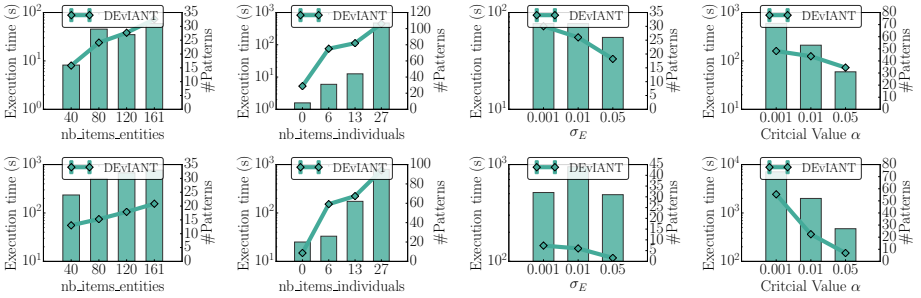


Fig. 5: Effectiveness of DEvIANT on Movielens when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

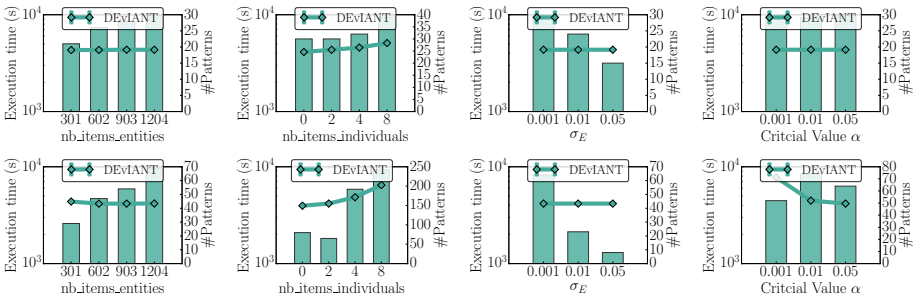


Fig. 6: Effectiveness of DEvIANT on Yelp when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

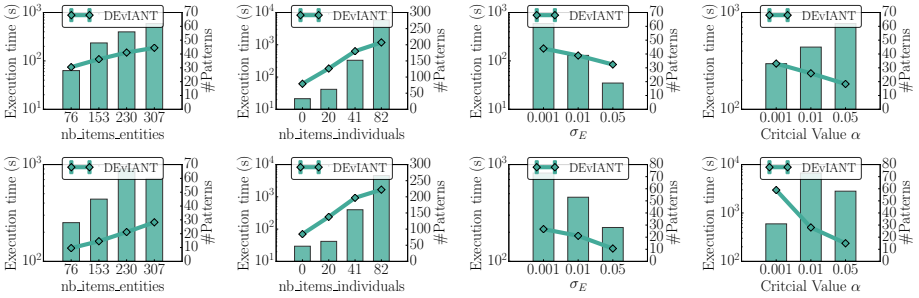


Fig. 7: Effectiveness of DEvIANT on CHUS when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

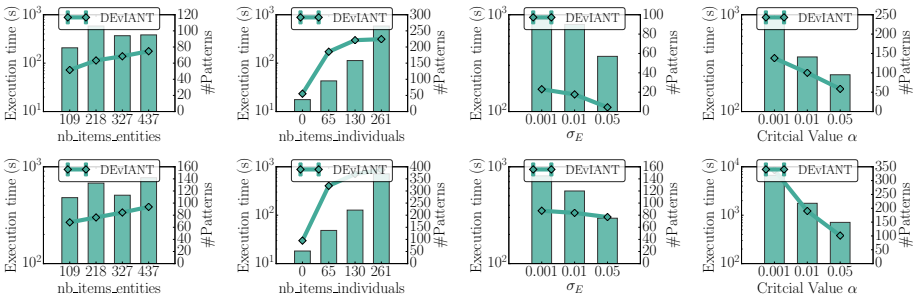


Fig. 8: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

D.2 Qualitative evaluation

Here, we report additional illustrative examples depicting the significant patterns discovered by DEvIANT when carried on the Eighth European Parliament (EPD8) dataset and Yelp dataset.

Table 7: Top-5 exceptional consensual/conflictual subjects among European Political Groups in the 8$^{\text{th}}$ EU parliament. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	p-$value$	IA
p_1	S&D	8.10 Revision of the Treaties and intergovernmental conferences	0.81	0.44	< 0.001	Conflict
p_2	*	2 Internal market, single market 6 External relations of the Union	0.27	0.54	< 0.001	Consensus
p_3	S&D	8.30 Treaties in general	0.81	0.55	< 0.001	Conflict
p_4	*	2 Internal market, single market, 4.15 Employment policy, act. combat unemployment	0.27	0.53	< 0.001	Consensus
p_5	ALDE	1.20.09 Protection of privacy and data protection 8 State and evolution of the Union	0.73	0.48	< 0.001	Conflict

Table 8: Top-10 exceptional consensual/conflictual subjects among countries' parlementarians in the 8$^{\text{th}}$ EU parliament. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	p-$value$	IA
p_1	Sweden	4 Economic, social and territorial cohesion 6.30 Development cooperation	0.3	0.84	<0.0001	Consensus
p_2	Finland	4 Economic, social and territorial cohesion 6.30 Development cooperation	0.36	0.87	<0.0001	Consensus
p_3	Finland	8.20.04 Pre-accession and partnership	0.36	0.75	<0.01	Consensus
p_4	Sweden	8.20 Enlargement of the Union	0.3	0.66	<0.0001	Consensus
p_5	Slovakia	1.10 Fundamental rights in the EU, Charter	0.48	0.13	<0.0001	Conflict
p_6	Malta	4.60.06 Consumers economic and legal interests	0.63	0.97	<0.0001	Consensus
p_7	Malta	2.10 Free movement of goods	0.63	0.34	<0.0001	Conflict
p_8	Latvia	4.60.06 Consumers economic and legal interests	0.42	0.69	<0.0001	Consensus
p_9	Luxembourg	1.20 Citizen's rights, 8 State and evolution of the Union	0.51	0.23	<0.01	Conflict
p_{10}	*	2 Internal market, single market 6 External relations of the Union	0.27	0.54	<0.001	Consensus

Table 9: Top-10 exceptional consensual/conflictual subjects among German national parties in the 8^{th} EU parliament. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	Sozialdemokratische Partei Deutschlands	1 European citizenship, 3 Community Policies	0.91	0.31	<0.0001	Conflict
p_2	*	3.70.11 Natural disasters, Solidarity Fund	0.38	0.93	<0.0001	Consensus
p_3	*	6.20.05 Multilateral economic and trade agreements and relations	0.38	0.85	<0.0001	Consensus
p_4	*	3.50 Research and technological development 4 Economic, social and territorial cohesion	0.38	0.78	<0.001	Consensus
p_5	*	3.30.03 Telecommunications, data transmission, telephone	0.38	0.02	<0.0001	Conflict
p_6	Liberal-Conservative Refomists	3.50.15 Intellectual property, copyright	0.91	0.57	<0.0001	Conflict
p_7	*	3.30.06 Information and communication tech. 4 Economic, social and territorial cohesion	0.38	0.04	<0.001	Conflict
p_8	DIE LINKE.	3.15 Fisheries policy	0.88	0.56	<0.0001	Conflict
p_9	*	3.50.20 Scientific and technological cooperation and agreements	0.38	0.7	<0.001	Consensus
p_{10}	*	3.30.05 Electronic and mobile communications, personal communications	0.38	0.07	<0.001	Conflict

Table 10: Top-10 exceptional consensual/conflictual places/categories/states among Yelp users. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	*	03 Automotive	0.14	-0.16	<0.0001	Conflict
p_2	*	10 Health & Medical	0.14	-0.14	<0.0001	Conflict
p_3	*	08 Financial Services	0.14	-0.11	<0.0001	Conflict
p_4	newcomer	09.38.07 Health Markets, 09.47 Juice Bars & Smoothies	0.14	-0.07	<0.01	Conflict
p_5	*	El Dorado Hills, California	0.14	0.35	<0.0001	Consensus
p_6	*	14 Local Services	0.14	-0.06	<0.0001	Conflict
p_7	*	04 Beauty & Spas	0.14	-0.06	<0.0001	Conflict
p_8	*	15 Mass Media	0.14	-0.05	<0.01	Conflict
p_9	*	11 Home Services'	0.14	-0.05	<0.0001	Conflict
p_{10}	*	Midlothian, Edinburgh	0.14	0.31	<0.0001	Consensus

E Appendix: Empirical DFDs

Here, we give an overview of the empirical distributions of Krippendorff's Alpha for 1000 draws from F_k equally likely to occur. Recall that F_k represents the subsets of the entire collection of entities of size k over which we define the random variable $\theta_k : F_k \to \mathbb{R}$. Thus, the distributions presented here illustrate the values observed on 1000 trials of θ_k. To illustrate the fact that the confidence intervals associated with θ_k (considering its distribution under the null hypothesis) are nested (when k grows, the confidence interval shrinks), we perform the experiments for various valuations of k.

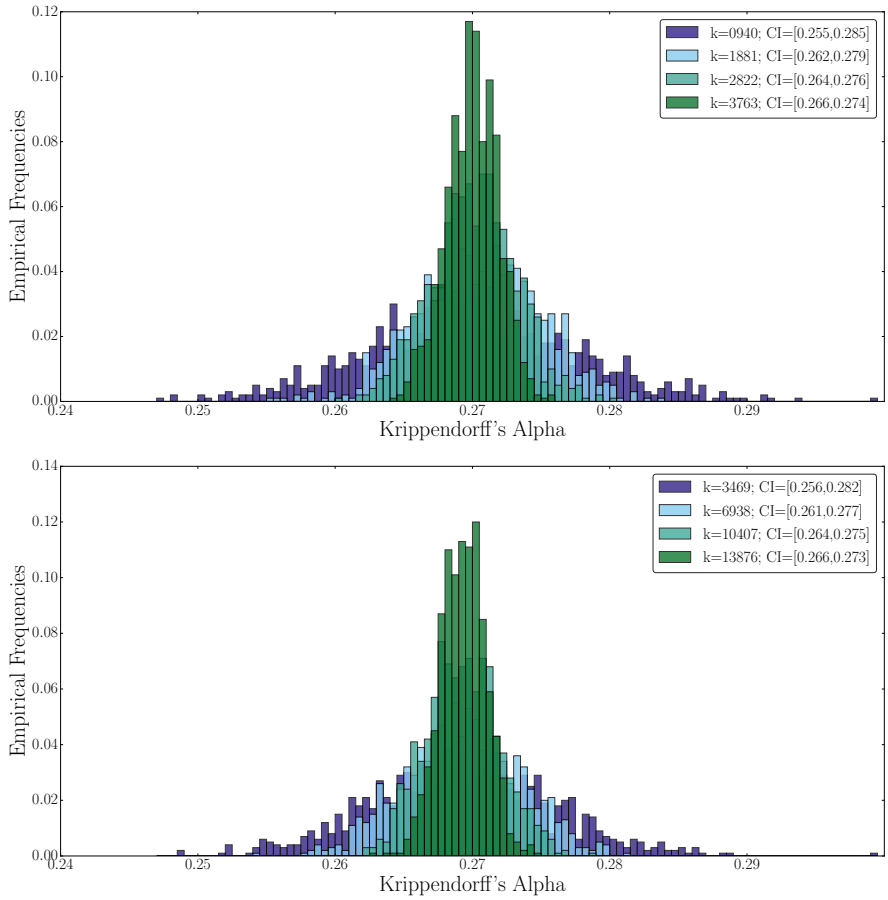


Fig. 9: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD). The top figure displays experiments on EPD8; the bottom figure displays experiments on CHUS (US House of representatives). We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

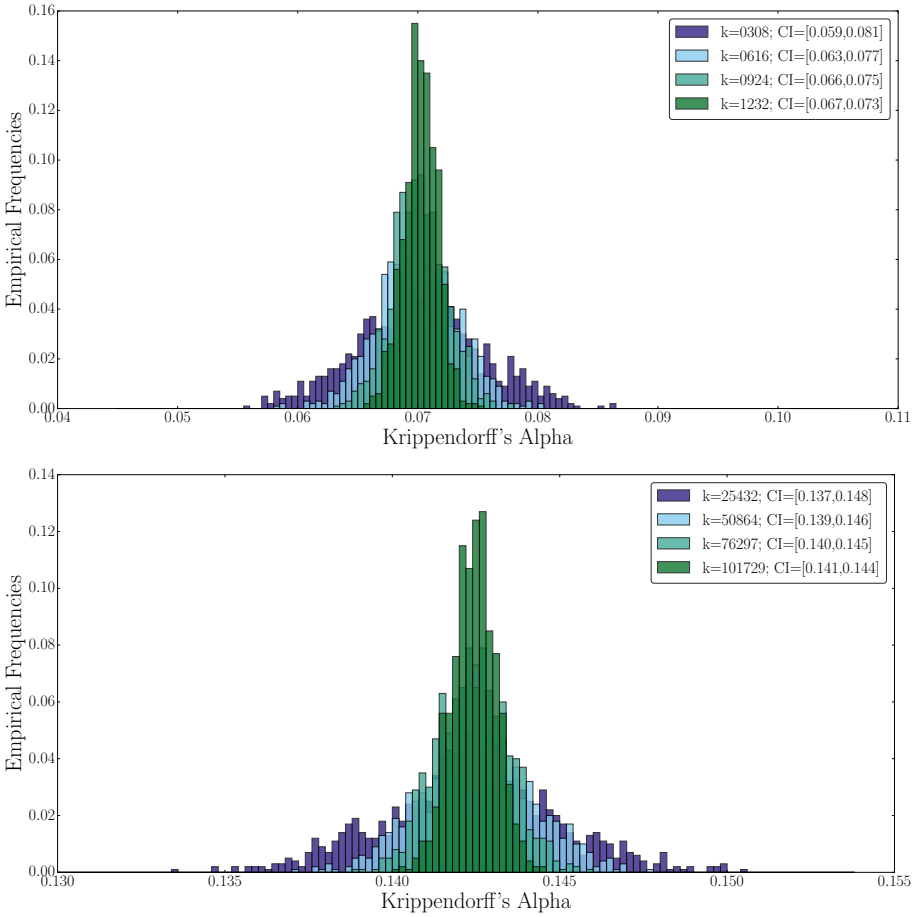


Fig. 10: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD). The top figure displays experiments on Movielens; the bottom figure displays experiments on Yelp. We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

Table 11: Symbol table

Symbol	Definition
G_E	A finite collection of records depicting entities
G_I	A finite collection of records depicting individuals
O	The domain of possible outcomes
o	Function returning the outcome of an individual over an entity
$\mathcal{B}$	$= \langle G_I, G_E, O, o\rangle$; A behavioral dataset
$\mathcal{A}$	$\mathcal{A}_E$ (resp. $\mathcal{A}_I$): Descriptive attributes of entities (resp. individuals)
$\mathcal{D}$	$\mathcal{D}_E$ (resp. $\mathcal{D}_I$): The description domain of contexts (resp. groups)
G_E^c	A subgroup of entities supporting a context $c \in \mathcal{D}_E$
G_I^g	A subgroup of individuals supporting a group $g \in \mathcal{D}_I$
g	$= g \in \mathcal{D}_I$; a description of a group of individuals characterizing $G_I^g \subseteq G_I$.
c	$= c \in \mathcal{D}_E$; a context characterizing a subset of entities $G_E^c \subseteq G_E$.
p	$= (g, c) \in \mathcal{P}$; The sought patterns.
P	$\subseteq \mathcal{P}$; The returned pattern set
$\sqsubseteq$	read "less restrictive than" is a partial order between descriptions in some descriptions space $\mathcal{D}$ ($\mathcal{D}_E$ or $\mathcal{D}_I$)
$\mathcal{B}^g$	The reduced behavioral dataset for individuals comprising G_I^g
A	Intra-group agreement measure - Krippendorff's Alpha
$A^g(G_E^c)$	Intra-group agreement of a group g over a context c
	We omit the exponent g in the notations and we assume that we have a group of individuals g in mind (we use $\mathcal{B}^g$)
D_{exp}	Expected disagreement (via marginal distribution) between individuals
D_{obs}	Observed disagreement between individuals
n	Number of entities in G_E, i.e., $\lvert G_E\rvert$
m	Number of all expressed outcomes
m^{o_1}	Number of expressed outcomes equal to o_1
m_e	Number of expressed outcomes for entity e (also denoted w_e)
$m_e^{o_1}$	Number of expressed outcomes equal to o_1 for entity e
$\delta_{o_1 o_2}$	Distance between two outcomes in O
DFD	Distribution of False discoveries
F_k	$F_k = \{S \subseteq G_E \text{ s.t. } \lvert S\rvert = k\}$
θ_k	Random variable $\theta_k : F_k \to \mathbb{R}$ with $S \mapsto A(S)$. Also $\theta_k = \frac{V_k}{W_k}$
v_e	Intra-group agreement (Krippendorff's Alpha) for one entity, given by: $m_e - \frac{1}{D_{\text{exp}}} \sum_{o_1,o_2 \in O^2} \delta_{o_1 o_2} \cdot \frac{m_e^{o_1} \cdot m_e^{o_2}}{(m_e - 1)}$
V_k	Random variable $V_k : F_k \to \mathbb{R}$ with $S \mapsto \frac{1}{k} \sum_{e \in S} v_e$
W_k	Random variable $W_k : F_k \to \mathbb{R}$ with $S \mapsto \frac{1}{k} \sum_{e \in S} w_e$
α	Critical value
$CI_k^{1-\alpha}$	The $1-\alpha$ confidence interval associated with the DFD of θ_k.
$\widehat{CI}_k^{1-\alpha}$	The $1-\alpha$ Taylor-approximated confidence interval of $CI_k^{1-\alpha}$.
$\widehat{CI}_{\text{bootstrap}}^{1-\alpha}$	The bootstrap confidence interval.
$LB(S, \sigma_E)$	Lower bound of A for any specialization of a subgroup having its size greater than σ_E
$UB(S, \sigma_E)$	Upper bound of A for any specialization of a subgroup having its size greater than σ_E
$OE(S, \sigma_E)$	$= [LB(S, \sigma_E), UB(S, \sigma_E)]$. Optimistic estimate region of A

References

1. S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. Lakshmanan, and R. H. Zamar. Exploring rated datasets with rating maps. WWW, 2017.
2. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data mining and knowledge discovery*, 5(3):213–246, 2001.
3. A. Belfodil, S. Cazalens, P. Lamarre, and M. Plantevit. Flash points: Discovering exceptional pairwise behaviors in vote or rating data. ECML/PKDD, 2017.
4. M. Boley, T. Horváth, A. Poigné, and S. Wrobel. Listing closed sets of strongly accessible set systems with applications to data mining. *Theoretical Computer Science*, 411(3):691–700, 2010.
5. T. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
6. M. Das, S. Amer-Yahia, G. Das, and C. Yu. Mri: Meaningful interpretations of collaborative ratings. *PVLDB*, 4(11):1063–1074, 2011.
7. T. De Bie. An information theoretic framework for data mining. KDD, 2011.
8. W. Duivesteijn, A. J. Feelders, and A. Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.
9. W. Duivesteijn and A. Knobbe. Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. ICDM, 2011.
10. W. Duivesteijn, A. J. Knobbe, A. Feelders, and M. van Leeuwen. Subgroup discovery meets bayesian networks - an exceptional model mining approach. ICDM, 2010.
11. O. J. Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.
12. F. Duris, J. Gazdarica, I. Gazdaricova, L. Strieskova, J. Budis, J. Turna, and T. Szemes. Mean and variance of ratios of proportions from categories of a multinomial distribution. *Journal of Statistical Distributions and Applications*, 5, 2018.
13. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
14. D. Eppstein and D. S. Hirschberg. Choosing subsets with maximum weighted average. *J. Algorithms*, 24(1):177–193, 1997.
15. B. Ganter and S. Kuznetsov. Pattern structures and their projections. ICCS,2001.
16. B. Ganter and R. Wille. *Formal concept analysis - mathematical foundations*. Springer, 1999.
17. S. Geisser. *Predictive Inference*, volume 55. CRC Press, 1993.
18. H. Grosskreutz, S. Rüping, and S. Wrobel. Tight optimistic estimates for fast subgroup discovery. ECML/PKDD, 2008.
19. W. Hämäläinen. Efficient discovery of the top-k optimal dependency rules with fisher's exact test of significance. In *ICDM*, pages 196–205. IEEE Computer Society, 2010.
20. W. Hämäläinen. Statapriori: an efficient algorithm for searching statistically significant association rules. *Knowl. Inf. Syst.*, 23(3):373–399, 2010.
21. W. Hämäläinen and G. I. Webb. A tutorial on statistically sound pattern discovery. *Data Min. Knowl. Discov.*, 33(2):325–377, 2019.
22. A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.
23. S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
24. M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*, 181(10):1989–2001, 2011.

25. G. M. P. van Kempen and L. J. van Vliet. Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry: The Journal of the International Society for Analytical Cytology*, 39(4):300–305, 2000.
26. M. Kendall, A. Stuart, and J. Ord. Kendall's advanced theory of statistics. v. 1: Distribution theory. 1994.
27. J. Komiyama, M. Ishihata, H. Arimura, T. Nishibayashi, and S.-i. Minato. Statistical emerging pattern mining with multiple testing correction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 897–906. ACM, 2017.
28. K. Krippendorff. Content analysis, an introduction to its methodology. 2004.
29. S. O. Kuznetsov. Learning of simple conceptual graphs from positive and negative examples. PKDD, 1999.
30. S. O. Kuznetsov and S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2-3):189–216, 2002.
31. M. van Leeuwen and A. J. Knobbe. Diverse subgroup set discovery. *Data Min. Knowl. Discov.*, 25(2):208–242, 2012.
32. F. Lemmerich, M. Becker, and M. Atzmueller. Generic pattern trees for exhaustive exceptional model mining. ECML/PKDD, 2012.
33. F. Lemmerich, M. Becker, P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Mining subgroups with exceptional transition behavior. KDD, 2016.
34. F. Llinares-López, M. Sugiyama, L. Papaxanthos, and K. Borgwardt. Fast and memory-efficient significant pattern mining via permutation testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 725–734. ACM, 2015.
35. S. Minato, T. Uno, K. Tsuda, A. Terada, and J. Sese. A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration. ECML/PKDD, 2014.
36. N. Mukhopadhyay. *Probability and statistical inference*. CRC Press, 2000.
37. L. Pellegrina and F. Vandin. Efficient mining of the most significant patterns with permutation testing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2070–2079. ACM, 2018.
38. J. M. Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.
39. R. Tarone. A modified bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.
40. A. Terada, D. duVerle, and K. Tsuda. Significant pattern mining with confounding variables. In *PAKDD (1)*, volume 9651 of *Lecture Notes in Computer Science*, pages 277–289. Springer, 2016.
41. A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32):12996–13001, 2013.
42. A. Terada, K. Tsuda, and J. Sese. Fast westfall-young permutation procedure for combinatorial regulation discovery. In G. Li, S. Kim, M. Hughes, G. J. McLachlan, H. Sun, X. Hu, H. W. Ressom, B. Liu, and M. N. Liebman, editors, *2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, December 18-21, 2013*, pages 153–158. IEEE Computer Society, 2013.
43. G. I. Webb. Discovering significant rules. In *KDD*, pages 434–443. ACM, 2006.
44. G. I. Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.
45. G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2-3):307–323, 2008.

46. P. H. Westfall, S. S. Young, et al. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993.
47. S. Wrobel. An algorithm for multi-relational discovery of subgroups. PKDD, 1997.