DEvIANT : Discovering significant exceptional (dis)agreement within groups

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

▶ **To cite this version:**

Adnene Belfodil, Wouter Duivesteijn, Marc Plantevit, Sylvie Cazalens, Philippe Lamarre. DEvIANT : Discovering significant exceptional (dis)agreement within groups. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019), Sep 2019, Würzburg, Germany. pp.3-20, 10.1007/978-3-030-46150-8_1 . hal-02961093

HAL Id: hal-02961093
https://hal.science/hal-02961093

Submitted on 8 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DEvIANT : Discovering significant exceptional (dis)agreement within groups

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 consider any type of data featuring individuals (e.g., parliamentarians, customers) performing observable actions (e.g., votes, ratings) on entities (e.g., legislative procedures, movies). In such data, we aim to find contexts (i.e. subgroup of entities) for which an exceptional (dis)agreement is observed among a group of individuals. To this end, we introduce the novel problem of discovering statistically significant exceptional contextual intra-group agreement patterns. To handle the data sparsity, we use the Krippendorff's Alpha measure to assess 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 to patterns for a computationally efficient significance assessment. We prove that these approximate CIs are nested along specialization of patterns. This makes it possible to incorporate pruning properties in the algorithm to early discard non-significant patterns. Empirical study on several datasets demonstrates the efficiency and the usefulness of DEvIANT.

1 Introduction

Consider data describing the organization and votes of the European Parliament (EP). Such 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 data offers interesting opportunities to study the agreement or disagreement of coherent subgroups, especially to highlight some unexpected ones. 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 ballot 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 votes on bills while

Amazon-like customers post ratings or reviews of products. A challenge with this approach when considering voting or rating data, however, is to effectively handle the absence of outcomes (sparsity) which is high in such datasets. For instance, in the European parliament data, MEPs votes on average on only a three-quarter of all sessions. These outcomes are not missing at random: special workgroups are often formed of MEPs who are tasked with properly studying a specific topic, and members of these workgroups are more likely to vote on their topic of study. 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.

In this paper, we introduce the problem of discovering significantly exceptional contextual intra-group agreement patterns, rooted in the Subgroup Discovey (SD) [42]/ Exceptional Model Mining (EMM) [8] framework. To tackle the data sparsity issue, we measure the agreement among groups with *Krippendorff's alpha* [27], a measure developed in the context of content analysis [28] which is well-known to handle missing outcomes elegantly. We develop a branch-and-bound algorithm to automatically find subgroups featuring statistically significant exceptional (dis)agreement among groups. This algorithm enables to early discard non significant subgroups by pruning unpromising branches of the search space. Fig. 1 illustrates this. Suppose, we are interested by subgroups of entities (e.g. voting sessions) whose sizes are greater than a support threshold σ. For a given subgroup of size $X \geq \sigma$, we gauge its exceptionality by its *p-value*, i.e. the probability of observing a quality (i.e. Krippendorff's alpha) for a random subset of entities is at least as extreme as the one observed for the subgroup. Hence avoiding to report subgroups which observe a high quality that appears due to chance only (i.e. accepting the null hypothesis). For this task to be achieved, we can estimate the empirical distribution of the quality of random

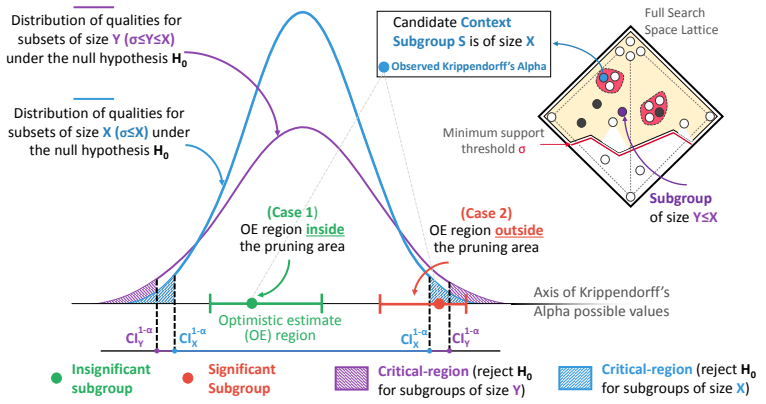


Fig. 1: Main DEvIANT properties for safe sub-search space pruning. A subgroup is reported as significant if its related Krippendorff's Alpha falls in the critical region of the corresponding empirical distribution of random subsets (DFD). When traversing the search space downward (decreasing support 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.

subsets (DFD: Distribution of False Discoveries, cf. [9,34]) and establish, for a certain critical value α, a confidence interval $CI_X^{1-\alpha}$ over the corresponding distribution under the null hypothesis. If the subgroup quality is outside $CI_X^{1-\alpha}$, this means that the subgroup is statistically significant ($p\text{-}value \leq \alpha$), otherwise the subgroup is a spurious finding. We prove by approximation that the confidence intervals are nested: $\sigma \leq Y \leq X \Rightarrow CI_X^{1-\alpha} \subseteq CI_Y^{1-\alpha}$. Moreover, we compute a tight optimistic estimate (OE) [18] to have a lower-bound and an upper bound of the quality of any specialization of a subgroup having its size greater than σ. Combining this two properties, if the OE region falls into the corresponding CI, we can safely prune large parts of the sub-search space that do not contain significant subgroups. In summary, the main contributions are:

- We introduce the novel problem of discovering statistically significant exceptional contextual intra-group agreement patterns (Section 3).
- We derive an analytical approximation of the confidence intervals associated to 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).
- 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 nested approximate CIs property to early discard non significant patterns.
- We report an empirical evaluation (Section 7) which studies the performance and the potential of the proposed approach.

2 Background and Related Work

Measuring Agreement. Several measures of agreement focus on two targets (Pearson's correlation coefficient, Spearman's ρ, Kendall's τ, Association) and 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 [4] is a seminal measure of agreement between two raters who classify items into a fixed number of mutually exclusive categories. The Fleiss κ [14] 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. The statistical assessment of patterns have received a considerable attention for a decade [41,20,21], especially for association rules [19,40,36]. Some works focused on the statistical significance of results in Subgroup Discovery/Exceptional Model Mining during enumeration [9,34] or a posteriori [10] for statistical validation of the found subgroups.

Voting and Rating data Analysis. In a previous work [2], we proposed a method to discover exceptional *inter-group agreement* in voting or rating data.

This method does not allow to discover intra-group agreement. Several works in the litterature addressed the problem of uncovering groups in rating datasets whose members exhibit an agreement or discord [6,38] or a specific rating distribution [1] (e.g. polarized, homogoneous) given upfront by the end-user. This is done by aggregating the ratings by using an arithmetic mean or a rating distribution. However, none of these methods allow to discover automatically exceptional (dis)agreement within groups. Moreover, these methods may output misleading hypotheses over the intra-group agreement. This is due to two main factors: aggregating ratings in a distribution is (i) highly affected by the data sparsity (e.g. two reviewees may have significantly different 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

The data we are interested in 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*). This type of data is called behavioral dataset (e.g. Tab. 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* $o : G_I \times G_E \to O$ *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 attributes potentially organized among a taxonomy. Subgroups (subsets) of G_I (resp. G_E) can be 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*, 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

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

(a) Entities

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

(b) Individuals

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

(c) Outcomes

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

$\mathcal{D}$ are partially ordered by a specialization operator denoted $\sqsubseteq$. A description d_1 is a specialization of d_2, denoted $d_1 \sqsubseteq d_2$, iff $d_2 \Rightarrow d_1$ from a logical point of view. It follows that $G^{d_2} \subseteq G^{d_1}$.

3.1 Krippendorff's Alpha

Krippendorff's Alpha (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 behavioral data with missing values; (iv) it takes into consideration the agreement expected by chance [28]. In its most general form, A is defined by:

$$A = 1 - \frac{D_o}{D_e} \tag{1}$$

where D_o (resp. D_e) is a measure of the disagreement observed (resp. by chance). 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 Eq. (1), we have $A(S) = 1 - \frac{D_o(S)}{D_e}$ for any $S \subseteq G_E$. Note that the measure takes into consideration only entities where at least two individuals expressed an outcome. We assume that the entities that do not fulfil this requirement are removed in preprocessing.

$$D_o(S) = \frac{1}{\sum_{e \in S} n_e} \sum_{c,k \in O^2} \delta_{ck} \cdot \sum_{e \in S} \frac{n_{ec} \cdot n_{ek}}{n_e - 1} \tag{2}$$

n_e is the number of expressed outcomes for the entity e and n_{ec} (resp. n_{ek}) represents the number of outcomes equal to c (resp. k) expressed for the entity e. δ_{ck} is a distance measure between outcomes, which can be defined according to the domain of the outcomes (e.g. δ_{ck} can correspond to the Kronecker delta for categorical outcomes or distance between ordinal values for ratings. Choices for the distance measure are discussed in [30]). We define below D_e that represents the disagreement expected by chance in Krippendorff's alpha:

$$D_e = \frac{1}{\sum_{e \in G_E} n_e \cdot (\sum_{e \in G_E} n_e - 1)} \sum_{c,k \in O^2} \delta_{ck} \cdot n_c \cdot n_k \tag{3}$$

With n_c (resp. n_k) the number of expressed outcomes equal to c (resp. k) observed in the entire behavioral data. This corresponds to the disagreement by chance observed on the overall marginal distribution of outcomes.

Example: Tab. 2 summarizes the behavioral data given in Tab. 1. The disagreement expected by chance is equal to (given: $nb(F) = 8$, $nb(A) = 6$): $D_e = \frac{1}{13\times14} \times ((8 \times 6) + (6 \times 8)) = 48/91$. If we want to evaluate the intra-agreement between the four individuals in the global context (considering all entities), we need to compute, first, the observed disagreement $D_o(G_E)$. Simply put, $D_o(G_E)$ is the weighted average of the two last lines by considering the quantities n_e as the weights: $D_o(G_E) = \frac{4}{14}$.Hence,for the global context, $A(G_E) = 1 - \frac{4}{14}/\frac{48}{91} = 0.46$. Consider the context $c = \langle themes \supseteq \{7.30 \text{ Judicial Coop.}\}\rangle$, 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: $D_o(G_E^c) = \frac{4}{7}$. Hence, the contextual intra-agreement is: $A(G_E^c) = 1 - \frac{4}{7}/\frac{48}{91} = -0.08$.

	[**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
n_e	3	2	2	2	3	2
$D_o(e)$	0	0	1	0	$\frac{2}{3}$	0

Table 2: Summarized Behavioral Data $D_o(e) = \sum_{c,k\in O^2} \delta_{ck} \frac{n_{ec} \cdot n_{ek}}{n_e \cdot (n_e - 1)}$

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

3.2 Mining Significant Patterns with Krippendorff's Alpha

Considering a selected group of individuals $g \in \mathcal{D}_I$, we are interested in finding contexts $c \in \mathcal{D}_E$ where the observed intra-agreement, denoted $A^g(c)$, *significantly* differs from the expected intra-agreement (i.e. that appear due to chance alone). In the same spirit as [9,34,41], we evaluate the quality of patterns by statistical significance of the contextual intra-agreement. This choice is motivated by: (i) the desire to not specify to the algorithm an arbitrary threshold on the distance from the overall intra-agreement observed (fixing the critical value α is more intuitive), (ii) the recommendations of Krippendorff [22] to provide a confidence interval on the alpha metric rather than a point-value.

In this work, we are interested in finding patterns of the form $(g, c) \in \mathcal{D}_I \times \mathcal{D}_E$ highlighting an exceptional intra-agreement between members of a group of individuals g over a context c. To perform such task, we formalize the 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].

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 corresponding to that the observed Krippendorff's alpha is generated by a *distribution of false discoveries* (cf. [9]). The goal is to find the pattern set $P \subseteq \mathcal{D}_I \times \mathcal{D}_E$ such that:
$$\forall (g, c) \in P : |G_I^g| \geq \sigma_I \wedge |G_E^c| \geq \sigma_E \text{ we have: } p\text{-}value^g(c) \leq \alpha$$

4 Exceptional Contexts: Evaluation and Pruning

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

4.1 Gauging Exceptionality of a Subgroup

To evaluate to what extent our findings are exceptional, we follow the significant pattern mining paradigm. That is, 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 a baseline finding. i.e. 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, c.f. [9]). We resolve to 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 a uniform sampling without replacement from the entire entities collection. The rationale behind using sampling without replacement is that the observed subgroup does not contain multiple instances of the same entity. 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 [34]

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), (2) and (3) as such:

$$A(S) = \frac{\sum_{e \in S} v_e}{\sum_{e \in S} w_e} \quad \text{with } w_e = n_e \text{ and } v_e = n_e - \frac{1}{D_e} \sum_{c,k \in O^2} \delta_{ck} \cdot \frac{n_{ec} \cdot n_{ek}}{(n_e - 1)} \tag{4}$$

Considering the null hypothesis H_0 and under the assumption that the underlying distribution of intra-agreements is a normal distribution[4] $\mathcal{N}(\mu_k, \sigma_k^2)$, one

[4] In the same line of reasoning of [7,35], 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 (*the one that takes 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]

can define $CI_k^{1-\alpha}$ by computing $\mu_k = E[\theta_k]$ and $\sigma_k^2 = \text{Var}[\theta_k]$. Doing so, requires either calculating estimator of such moments empirically by drawing a large number r of uniformly generated samples from F_k or deriving analytically 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 being estimated empirically 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, on one hand, due to the problem setting, establishing the confidence interval empirically is computationally expensive since it need to be calculated for each enumerated context which can become quickly unfeasible even for relatively small behavioral datasets. In the other hand, deriving analytically a computationally efficient form of $E(\theta_k)$ is notoriously difficult, given that $E[\theta_k] = \frac{1}{\binom{n}{k}}\sum_{S\in F_k}\frac{\sum_{e\in S}v_e}{\sum_{e\in S}w_e}$ and $\text{Var}[\theta_k] = \frac{1}{\binom{n}{k}}\sum_{S\in F_k}\left(\frac{\sum_{e\in S}v_e}{\sum_{e\in S}w_e} - E[\theta_k]\right)^2$.

Since θ_k can be seen as a weighted arithmetic mean, one can model the random variable θ_k as the ratio $\frac{V_k}{W_k}$. With V_k, W_k 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 [11] and [25, p.351], since no easy analytic expression of $E[\theta_k]$ and $\text{Var}[\theta_k]$ can be derived. For the sake of brevity, the detailed computation of the formulas presented next are omitted. For more details, please refer to Appendix A.

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

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

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

4.2 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. We leverage results of Eppstein and Hirschberg [13] who propose an elegant *linear algorithm* `Random-SMWA`[5] to find subsets with maximum weighted average. Remind that A can be seen as a weighted average from Eq. (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 and given σ_E a user defined support threshold on the minimum allowed size of context extents, k is fixed to $|S| - \sigma_E$. The obtained subset correspond to the smallest allowed subset having its support $\geq \sigma_E$ maximizing the weighted average quantity A. The `Random-SMWA` algorithm can be tweaked[6] 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 the $CI_{k_2}^{1-\alpha}$. 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 property allows to prune sub-search space related to a context c when traversing the search space downward if the optimistic estimate $OE(G_E^c, \sigma_E) \subseteq CI_{|G_E^c|}^{1-\alpha}$.

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

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

[6] 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:*

$$\text{if } \sigma_E \geq C^{\alpha} = \frac{4n\beta_w^2}{z^2_{1-\frac{\alpha}{2}}(\beta_v+\beta_w)+4\beta_w^2} \text{ then:}$$

$$\forall k_1, k_2 \in \mathbb{N} : \sigma_E \leq k_1 \leq k_2 \Rightarrow \widehat{CI}^{1-\alpha}_{k_2} \subseteq \widehat{CI}^{1-\alpha}_{k_1}$$

Combining Properties (1), (2) and (3), we formalize the pruning region property which answers to: *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 $\sigma_E \geq C^{\alpha}$, 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:*

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

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 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 a 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* [12] 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 individual g. Following the bootstrapping scheme proposed by Krippendorff [29,22,43] , 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 $n_e \cdot (n_e - 1)$ pairs of outcomes according to the distribution of the observed pairs of outcomes; (3) compute the disagreement observed and calculate the Krippendorff alpha quantity on the resulting resample. This process, repeated b times leads to a vector of bootstrap estimates (sorted in ascending order) $\hat{B} = [\hat{A}_1, ..., \hat{A}_b]$. 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. $\text{CI}^{1-\alpha}_{\text{bootstrap}} = [\hat{B}_{\frac{\alpha}{2}\cdot b}, \hat{B}_{(1-\frac{\alpha}{2})\cdot b}]$. 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

Subgroup Enumeration. In order to detect exceptional contextual intra-group agreement patterns, we need to enumerate candidate $p = (g, c) \in (\mathcal{D}_I, \mathcal{D}_E)$. For this task, different enumeration algorithms exist in the literature ranging from heuristic (e.g. beam-search [32]) to exhaustive techniques (e.g. GP-growth [33]). In this paper, we choose to 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 our previous work [2], we apply EnumCC (Enumerate Closed Candidates) [2] to enumerate subgroups g (resp. c) in $\mathcal{D}_I$ (resp. $\mathcal{D}_E$). EnumCC goes in the same line of CloseByOne algorithm [31]. Given G a collection of records (which can be either G_E or G_I), EnumCC traverses the search space in a DFS fashion and enumerates once and only once all the closed descriptions that fulfill the minimum support constraint σ. For more details, see Appendix B.

DEvIANT (Algorithm 1) 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 individual. Next, the corresponding behavioral dataset $\mathcal{B}^g$ is established by reducing the original behavioral dataset $\mathcal{B}$ to elements concerning solely the individuals comprising G_I^g. Subsequently, the bootstrap confidence interval $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$ is calculated.

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

Inputs : $\mathcal{B} = \langle G_I, G_E, O, o\rangle$ is a behavioral dataset,
σ_E (resp. σ_I) minimum support threshold of a context (resp. group),
α is a critical significance value (fixed to 0.05 in default setting).

Output: P is the set of exceptional intra-group agreement patterns

$P \leftarrow \{\}$
foreach $(g, G_I^g, cont_g) \in \mathrm{EnumCC}(G_I, *, \sigma_I, 0, True)$ **do**
  $G_E(g) = \{e \in E \text{ s.t. } n_e^g \geq e\}$
  $\mathcal{B}^g = \langle G_E(g), G_I^g, O, o\rangle$
  $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}} = [\hat{B}_{\frac{\alpha}{2}\cdot b}, \hat{B}_{(1-\frac{\alpha}{2})\cdot b}]$ ▷ 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

Before searching for exceptional contexts, the minimum context support threshold σ_E is adjusted to $C^\alpha(g)$ (see Prop.3) if it is lower than $C^\alpha(g)$. Note that $C^\alpha(g)$ is, in practice, much smaller than σ_E. Still, we keep this correction for algorithm soundness. Next, contexts are enumerated by EnumCC and for each candidate context c, the algorithm evaluates the optimistic estimate interval $OE(G_E^c)$ (see Prop. 2). According to Corollary 1, if $OE(G_E^c, \sigma_E^g)$ is inside $\text{MCI}^{1-\alpha}_{|G_E^c|}$ the sub-search space under c can be pruned. Otherwise, $A^g(G_E^c)$ is computed and evaluated against $\text{MCI}^{1-\alpha}_{|G_E^c|}$. If it is outside $\text{MCI}^{1-\alpha}_{|G_E^c|}$, this means that (g, c) is significant and should be kept in the final result set P. Eventually, to reduce the number of reported patterns, we choose to 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: $\forall p' = (g', c') \in \mathcal{P} \backslash P : p\text{-}value^{g'}(c') \leq \alpha \Rightarrow \exists p = (g, c) \in P$ 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 her analysis to obtain more fine-grained results by re-launching the algorithm starting from a specific context or group.

7 Empirical Evaluation

We report on both quantitative and qualitative experiments over the implemented algorithms. For reproducibility purposes, the source code and the data are made available in our companion page[7]. The following experiments aim to answer the following questions: **($\mathbf{Q}_1$)** How well the Taylor approximate CI approaches the empirical CI? **($\mathbf{Q}_2$)** How efficient is the Taylor approximate CI and the pruning properties? **($\mathbf{Q}_3$)** Does DEvIANT provide interpretable patterns?
Datasets: experiments were carried on four real-world behavioral datasets (see Tab. 3). Two voting datasets (EPD8 and CHUS) and two rating datasets (Movielens and Yelp). Each dataset features entities and individuals that are described by categorical (C), numerical (N) attributes, or categorical attributes augmented with a taxonomy (H). We report also the equivalent number of items (in an itemset language) corresponding to the descriptive attributes (ordinal scaling [16]).

	$\lvert G_E\rvert$	$\mathcal{A}_E$ (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%	$\simeq 10^{-7}$
CHUS[9]	17350	$1H + 2N$ (307)	1373	$2C$ (261)	$3M$ (C)	31.2%	$\simeq 10^{-6}$
Movielens[10]	1681	$1H + 1N$ (161)	943	$3C$ (27)	$100K$ (O)	06.3%	$\simeq 0.06$
Yelp[11]	$127K$	$1H + 1C$ (851)	$1M$	$3C$ (6)	$4.15M$ (O)	0.003%	$\simeq 1.73$

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

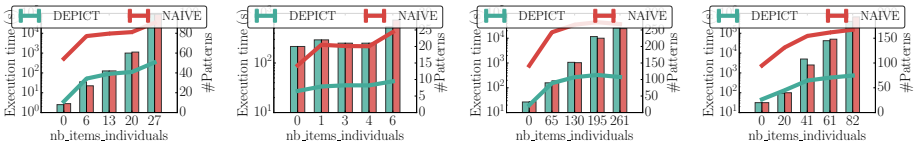
[7] https://github.com/Adnene93/Deviant
[8] Eight European Parliament Voting Dataset.
[9] 102^{nd}-115^{th} congresses of the US House of representatives (Period: 1991-2015).
[10] Movie review dataset - https://grouplens.org/datasets/movielens/100k/
[11] Social network dataset - https://www.yelp.com/dataset/challenge

Q_1- First, we evaluate to what extent the confidence interval computed empirically approximates the confidence interval computed by Taylor approximations. For this task, we run 1000 experiments for different subsets size k uniformly distributed in $[1, n = |G_E|]$. For each k, we compute the corresponding Taylor approximate $\widehat{CI}_k^{1-\alpha} = [a^T, b^T]$ and empirical confidence interval $\text{ECI}_k^{1-\alpha} = [a^E, b^E]$. The latter CI is calculated by running 10^4 samples of size k from G_E, followed by the computation of the observed A on each sample which are used to estimates the moments of the empirical distribution required for the establishment $\text{ECI}_k^{1-\alpha}$. Once both CIs are computed, we measure the distance between them by using the Jaccard index, i.e. $\text{dist}(\text{ECI}_k^{1-\alpha}, \widehat{\text{CI}}_k^{1-\alpha}) = 1 - \frac{\left(\min(b^E, b^T) - \max(a^E, a^T)\right)}{\left(\max(b^E, b^T) - \min(a^E, a^T)\right)}$. We report in Tab. 4, the average μ_{err} and the standard deviation σ_{err} of the observed distances (coverage error) over the 1000 experiments. We notice that the difference between the analytic Taylor approximation and the empirical approximation is negligible (μ_{err} is less than 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- In order to evaluate how efficient are the proposed properties (**(i)** Taylor approximate CI, **(ii)** optimistic estimates and **(iii)** nested approximates CIs), we choose to compare DEvIANT against 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 estimating empirically the confidence interval by successive random

$\mathcal{B}$	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	0.008	0.007

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



(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 outputted patterns. Parameters: $\sigma_E = \sigma_I = 1\%$ and $\alpha = 0.05$.

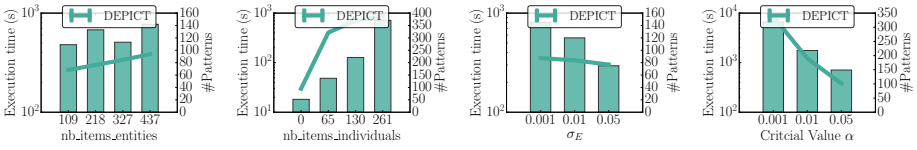


Fig. 3: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha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

trials from F_k. For this study, we choose to run both algorithms while disabling the bootstrap $\text{CI}^{1-\alpha}_{\text{bootstrap}}$ computation since the overhead induced by the computation of $\text{CI}^{1-\alpha}_{\text{bootstrap}}$ is the same for both algorithms. We vary the size of the descriptions space related to groups of individuals $\mathcal{D}_I$ while considering the whole description space of entities. Results of this experiment are reported in Fig.2. We observe that DEvIANT outperforms `Naive` in terms of runtime by nearly two orders of magnitude while outputting the same number of the desired patterns.

Furthermore, we report in Fig.3 the performance of DEvIANT in terms of runtime and the number of outputted patterns. We observe that when varying the descriptions space sizes, DEvIANT requires more time to finish. It is worth mentioning that the size of individuals search space $\mathcal{D}_I$ substantially affects the runtime of DEvIANT. This is mainly due to the fact that larger $\mathcal{D}_I$ leads to more candidate group of individuals g which require: (i) DEvIANT to generate the bootstrapping confidence interval $\text{CI}^{1-\alpha}_{\text{bootstrap}}$ and (ii) to mine for exceptional contexts c concerning the candidate group g. Finally, we observe that when α decreases, the execution time required for DEvIANT to finish increases while returning more patterns. This may at first, seem counter-intuitive since less patterns are naturally considered significant when alpha decreases, but this is supported by the fact that DEvIANT considers in the resulting pattern set only the most general patterns. Hence, when α decreases, DEvIANT goes deeper in the contexts search space, implying thus much more candidate patterns to be tested and thus a larger results set. The same conclusions can be drawn as well from experiments performed on Yelp, Movielens and CHUS (see Appendix. C).

$\mathbf{Q}_3$- We illustrate a number of examples of the outputted exceptional contextual intra-group agreement patterns in the benchmark datasets. Table 5 reports the exceptional contexts observed among the republicans party during the 115^{th} congress. For instance, Pattern p_1 illustrated in Fig. 4, highlights a collection of voting sessions addressing Government and Administrative issues where a clear polarization is observed between two clusters of House republicans. Notable roll call vote of this context in which a significant disagreement was observed between republicans is "**House Vote 417**"[12] which was closely watched by the media[13].

Table 6 depicts some patterns returned by DEvIANT when carried on Movielens Datasets. For instance, pattern p_2 reports that "Middle-aged Men" group observe a significantly higher intra-group agreement compared to the overall intra-group agreement for movies labeled with both adventure and musical genres (e.g. The Wizard of Oz (1939)).

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	p-$value$	IA
p_1	Republicans	20.11 Government Branch Relations, Admin. Issues, and Constitutional Reforms	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

Table 5: Exceptional consensual/conflictual subjects among Republicans Party representatives in the 115^{th} congress of the US House of Representatives. $\alpha = 0.01$

[12] https://projects.propublica.org/represent/votes/115/house/1/417

[13] Washington Post:https://wapo.st/2W32I9c; Reuters:https://reut.rs/2TF0dgV

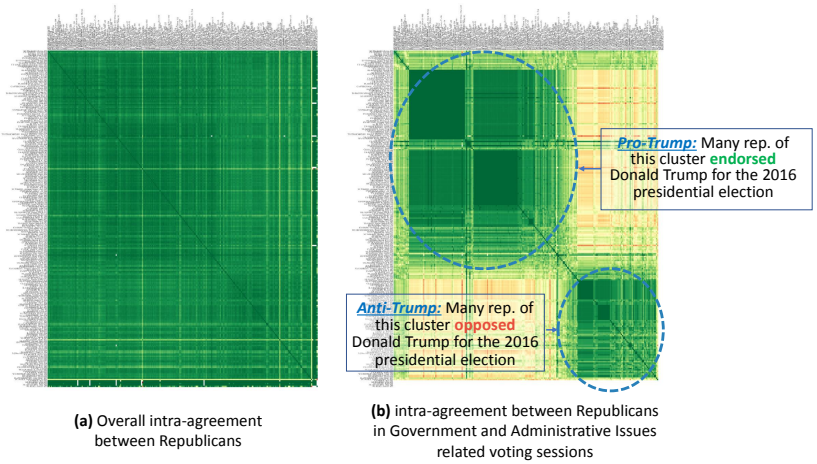


Fig. 4: Illustrating Pattern 1 from Tab. 5 with a similarity matrix between Republicans. Each cell represents the ratio of voting sessions in which both Rep. agreed. A green cell reports a strong agreement whereas a red cell highlights a strong disagreement.

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

Table 6: Top3-Exceptionally consensual/conflictual movies genres between Movielens' raters, α=0.01. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

8 Conclusion and Future Directions

In this paper we introduced the novel problem of discovering statistically significant exceptional contextual intra-group agreement patterns. We devised a branch-and-bound algorithm, named DEvIANT, which efficiently search for the desired patterns while leveraging closure operators, approximate confidence intervals (CIs), tight optimistic estimates on Krippendorff's Alpha measure and the property of nested CIs. The empirical experiments demonstrated both the efficiency and the usefulness of DEvIANT over multiple behavioral datasets relevant to various domains ranging from political analysis to rating data analysis. In future research, we plan (i) to incorporate FDR (False Discovery Rate) control to tackle the multiple comparison problem [21] (ii) to investigate exceptional intra-group agreement compared to the one observed between all individuals over the same context and (iii) to integrate the option of choosing which kind of exceptional consensus the end-user want, i.e. is the exceptional consensus is observed because the group members "liked/voted for" (or "disliked/voted against") the context related entities? All this being done within the perspective to provide a comprehensive framework and tool[14]) for behavioral data analysis alongside exceptional inter-group agreement pattern discovery implemented in [2].

[14]A prototype is available online in http://contentcheck.liris.cnrs.fr

References

1. S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. Lakshmanan, and R. H. Zamar. Exploring rated datasets with rating maps. In *WWW*, pages 1411–1419. International World Wide Web Conferences Steering Committee, 2017.
2. A. Belfodil, S. Cazalens, P. Lamarre, and M. Plantevit. Flash points: Discovering exceptional pairwise behaviors in vote or rating data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 442–458. Springer, 2017.
3. 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.
4. J. Cohen. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46, 1960.
5. T. M. Cover and J. A. 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. In *KDD*, pages 564–572. ACM, 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. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 151–160. IEEE, 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. 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(1):2, 2018.
12. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
13. D. Eppstein and D. S. Hirschberg. Choosing subsets with maximum weighted average. *J. Algorithms*, 24(1):177–193, 1997.
14. J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
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. In *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 440–456. Springer, 2008.
19. W. Hämäläinen. Statapriori: an efficient algorithm for searching statistically significant association rules. *Knowl. Inf. Syst.*, 23(3):373–399, 2010.
20. W. Hämäläinen and G. I. Webb. Statistically sound pattern discovery. In *KDD*, page 1976. ACM, 2014.
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. 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.
24. 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.
25. M. Kendall, A. Stuart, and J. Ord. Kendall's advanced theory of statistics. v. 1: Distribution theory. 1994.
26. A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. The big data bootstrap. In *ICML*. icml.cc / Omnipress, 2012.
27. K. Krippendorff. Estimating the reliability, systemic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.
28. K. Krippendorff. Content analysis, an introduction to its methodology. 2004.
29. K. Krippendorff. Bootstrapping distributions for krippendorff's alpha. *Available at web. asc. upenn. edu/usr/krippendorff/boot. c-Alpha. pdf*, 2006.
30. K. Krippendorff. Computing krippendorff's alpha-reliability. 2011.
31. 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.
32. M. van Leeuwen and A. J. Knobbe. Diverse subgroup set discovery. *Data Min. Knowl. Discov.*, 25(2):208–242, 2012.
33. F. Lemmerich, M. Becker, and M. Atzmueller. Generic pattern trees for exhaustive exceptional model mining. In *ECML/PKDD*, pages 277–292, 2012.
34. F. Lemmerich, M. Becker, P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Mining subgroups with exceptional transition behavior. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 965–974. ACM, 2016.
35. J. Lijffijt, B. Kang, W. Duivesteijn, K. Puolamäki, E. Oikarinen, and T. De Bie. Subjectively interesting subgroup discovery on real-valued targets. In *ICDE*, pages 1352–1355. IEEE Computer Society, 2018.
36. 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. In *ECML/PKDD (2)*, volume 8725 of *Lecture Notes in Computer Science*, pages 422–436. Springer, 2014.
37. N. Mukhopadhyay. *Probability and statistical inference*. CRC Press, 2000.
38. B. Omidvar-Tehrani, S. Amer-Yahia, P.-F. Dutot, and D. Trystram. Multi-objective group discovery on the social web. In *ECMLPKDD*, 2016.
39. J. M. Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.
40. 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.
41. G. I. Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.
42. S. Wrobel. An algorithm for multi-relational discovery of subgroups. PKDD, 1997.
43. A. Zapf, S. Castell, L. Morawietz, and A. Karch. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(1):93, 2016.

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 $F_k = \{S \in G_E \; s.t. \; |S| = k\}$. F_k is then the set of possible outcomes which are equally likely to occur under the null hypothesis H_0. G_E contains n records (i.e. $|G_E| = n$). Each record $e \in G_E$ is associated to a value v_e and w_e. θ_k can be expressed as a ratio of two random variable $\frac{V_k}{W_k}$. With V_k, W_k 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 (a concise lecture note[15] follows the same reasoning and explains the derivations) :

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

with 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:

$$E[f(x,y)] \approx f(\lambda) + \frac{1}{2}\left[f''_{xx}(\lambda)\mathrm{Var}[X] + f''_{xy}(\lambda)\mathrm{Cov}[X,Y] + f''_{yy}(\lambda)\mathrm{Var}[Y]\right] \tag{8}$$

Given that $f(x, y) = \frac{x}{y}$ and using the fact that $E[X - \mu_x] = 0$ (This 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}).

$$E[\theta_k] = E[\frac{V_k}{W_k}] = E[f(V_k, W_k)] \approx \frac{\mu_{V_k}}{\mu_{W_k}} - \frac{\mathrm{Cov}[V_k, W_k]}{\mu^2_{W_k}} + \frac{\mathrm{Var}[W_k]\mu_{V_k}}{\mu^3_{V_k}} \tag{9}$$

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 $n = |G_E|$.

$$E[V_k] = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \frac{1}{k}\sum_{e\in S} v_e = \frac{1}{n}\sum_{e\in G_E} v_e = \mu_v \tag{10}$$

$$\begin{aligned} \mathrm{Var}[V_k] &= \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - E[V_k]\right)^2 = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - \mu_v\right)^2 \\ &= \frac{1}{k}\left(\frac{n}{n-1}\left(\mu_{v^2} - \mu_v^2\right)\right) - \frac{1}{n-1}\left(\mu_v^2 - \mu_{v^2}\right) \text{ with } \mu_{v^2} = \frac{1}{n}\sum_{e\in G_E} v_e^2 \end{aligned} \tag{11}$$

[15] see http://www.stat.cmu.edu/~hseltman/files/ratio.pdf

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

$$E[W_k] = \frac{1}{n} \sum_{e \in G_E} w_e = \mu_w \tag{12}$$

$$\begin{aligned} \mathrm{Var}[W_k] &= \frac{1}{\binom{n}{k}} \sum_{S \in F_k} \left(\frac{1}{k} \sum_{e \in S} w_e - E[W_k] \right)^2 \\ &= \frac{1}{k} \left(\frac{n}{n-1} \left(\mu_{w^2} - \mu_w^2 \right) \right) - \frac{1}{n-1} \left(\mu_w^2 - \mu_{w^2} \right) \text{ with } \mu_{w^2} = \frac{1}{n} \sum_{e \in G_E} w_e^2 \end{aligned} \tag{13}$$

We derive now 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:

$$\begin{aligned} \mathrm{Cov}[V_k, W_k] &= \frac{1}{\binom{n}{k}} \sum_{S \in F_k} \left(\frac{1}{k} \sum_{e \in S} v_e - E[V_k] \right) \left(\frac{1}{k} \sum_{e \in S} w_e - E[W_k] \right) \\ &= \frac{1}{k} \left(\frac{n}{n-1} \left(\mu_{vw} - \mu_v \mu_w \right) \right) - \frac{1}{n-1} \left(\mu_v \mu_w - \mu_{vw} \right) \\ &\text{with } \mu_{vw} = \frac{1}{n} \sum_{e \in G_E} w_e v_e \end{aligned} \tag{14}$$

Using equations *(10)*, *(11)*, *(12)*, *(13)*, *(14)*, we derive the approximation of $E[\theta_k]$ after simplifications of *(9)*.

$$E[\theta_k] \approx \widehat{E}[\theta_k] = \left(\frac{n}{k} - 1 \right) \frac{\mu_v}{\mu_w} \beta_w + \frac{\mu_v}{\mu_w} \text{ with } \beta_w = \frac{1}{n-1} \left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \tag{15}$$

The same reasoning applies to approximate $\mathrm{Var}[\theta_k]$ using Taylor expansion. 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. Same observation have been made by [24,11] and [25, p.351] to approximate the variance for a ratio random variable. We obtain:

$$\mathrm{Var}[\theta_k] = \mathrm{Var}[f(V_k, W_k)] \approx \frac{\mathrm{Var}[V_k]}{\mu_{W_k}^2} - 2 \frac{\mu_{V_k} \mathrm{Cov}[V_k, W_k]}{\mu_{W_k}^3} + \frac{\mu_{V_k}^2 \mathrm{Var}[W_k]}{\mu_{W_k}^4} \tag{16}$$

After simplifications and by using the same line of reasoning when deriving the expected value approximation reported in equation (15), we obtain:

$$\begin{aligned} \mathrm{Var}[\theta_k] &\approx \widehat{\mathrm{Var}}[\theta_k] = \left(\frac{n}{k} - 1 \right) \frac{\mu_v^2}{\mu_w^2} \left(\beta_v + \beta_w \right) \\ &\text{with } \beta_w = \frac{1}{n-1} \left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \text{ and } \beta_v = \frac{1}{n-1} \left(\frac{\mu_{v^2}}{\mu_v^2} - \frac{\mu_{vw}}{\mu_v \mu_w} \right) \end{aligned} \tag{17}$$

We denote by $\widehat{CI}_k^{1-\alpha}$ the approximate confidence interval calculated using the approximations of the expected value $\widehat{E}[\theta_k]$ (15) and the variance $\widehat{\text{Var}}[\theta_k]$ (17) that is:

$$\widehat{CI}_k^{1-\alpha} = \left[\widehat{E}[\theta_k] - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}, \widehat{E}[\theta_k] + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_k]}\right] \tag{18}$$

It is worth mentioning that, the complexity of the computation of the approximate confidence interval (18) is linear to the size n. □

Proof (Proposition 2). To alleviate 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 [13].
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 reduction to absurdity. We denote by $S'_{\text{max}} \subset S'$ (resp. $S_{\text{max}} \subset S$) 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, namely S'_{max}, in S that observes a greater weighted average A than the actual optimal subset S_{max}, which is absurd.

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

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

$$k_1, k_2 \in \mathbb{N} : \text{ if } k_1 \leq k_2 \Rightarrow \widehat{\text{Var}}[\theta_{k_1}] \geq \widehat{\text{Var}}[\theta_{k_2}] \tag{19}$$

From (17), $\widehat{\text{Var}}[\theta_k] = \left(\frac{n}{k} - 1\right)\frac{\mu_v^2}{\mu_w^2}(\beta_v + \beta_w)$. Given that $\frac{n}{k} - 1$ is a decreasing function w.r.t. k, proving (19) 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 (16) is positive. This can be proved by a direct application of the Covariance inequality [37, p, 149],which itself is an application of the Cauchy-Schwarz inequality [39]. Since $\beta_v + \beta_w$ is of the same sign of (17), we have $\beta_v + \beta_w \geq 0$. For the sake of a self-contained proof of this assertion below:

We have, from (16) is of the same sign of:

$$\frac{\text{Var}[V_k]}{\mu_{V_k}^2} - 2\frac{\text{Cov}[V_k, W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\text{Var}[W_k]}{\mu_{W_k}^2} \tag{20}$$

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 we have the quantity (20) is greater than:

$$\frac{\sigma^2[V_k]}{\mu^2_{V_k}} - 2\frac{\sigma[V_k]\sigma[W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\sigma^2[W_k]}{\mu^2_{W_k}} \tag{21}$$

This quantity can be rewritten as such:

$$\frac{\sigma[V_k]}{\mu_{V_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) - \frac{\sigma[W_k]}{\mu_{W_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) \tag{22}$$

This quantity is clearly positive since (22) is equal to $\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right)^2 \geq 0$. Hence $\beta_v + \beta_w \geq 0$ which confirms that the variance is deceasing w.r.t. increasing size k as stated in equation (19).

Recall that, we want to ensure by approximation 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 find a lower bound for σ_E such that:

$$z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_{k_1}]} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}[\theta_{k_2}]} \geq \left|\widehat{E}[\theta_{k_1}] - \widehat{E}[\theta_{k_2}]\right| \tag{23}$$

By using the definitions of $\widehat{\text{Var}}[\theta_k]$ and $\widehat{E}[\theta_k]$ from equations (15) and (17), the inequality (23) can be rewritten as such:

$$\left(\sqrt{\frac{n}{k_1} - 1} + \sqrt{\frac{n}{k_2} - 1}\right) \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta^2_w}} \tag{24}$$

And since $\sigma_E \leq k_1 \leq k_2$:

$$2\sqrt{\frac{n}{\sigma_E} - 1} \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta^2_w}} \tag{25}$$

After simplifications, we obtain:

$$\sigma_E \geq C^\alpha = \frac{4n\beta^2_w}{z^2_{1-\frac{\alpha}{2}}(\beta_v + \beta_w) + 4\beta^2_w} \tag{26}$$

□

Proof (Corollary 1). The proof is a 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:

$$A(G^d_E) \in OE(G^c_E, \sigma_E) \tag{27}$$

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

$$CI^{1-\alpha}_{|G^c_E|} \subseteq \widehat{CI}^{1-\alpha}_{|G^d_E|} \tag{28}$$

From (27) and (28) and given that $OE(G^c_E, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G^c_E|}$, it follows that $A(G^d_E) \in OE(G^c_E, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G^c_E|} \subseteq \widehat{CI}^{1-\alpha}_{|G^d_E|}$ hence *p-value(d)* $> \alpha$. □

B Appendix: Enumeration Algorithm

Given a collection of records G which descriptive attributes are $\mathcal{A} = \{a_1, ..., a_m\}$ 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 conjunction 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. G^d denotes the set of records of G covered by the description d.

EnumCC algorithm enumerates once and only once all closed descriptions whose associated subgroups fullfill the minimum support constraint σ_E. The algorithm follows the same reasoning of most common SD algorithms and goes in the same line of CloseByOne Algorithm (CbO) [31] and Divide-And-Conquer Algorithm [3]. 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 $i \neq j$ intacts, and refines only the j^{th} condition. If the condition is related to a numerical attribute a left or a right minimal change is performed [23]. 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 [2]. In a nutshell, for each parameter description d, EnumCC starts by assessing if the subgroup G^d is valid ($|G^d| \geq \sigma$) and. In this case, the closed description `closure_d` is computed and returned only if the canonicity test is passed (cf. [16, p.66-68]). `closure_d` corresponds to the tightest description of G^d (maximum in the sense of partial order $\sqsubseteq$ ordering descriptions in $\mathcal{D}$) which is the

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

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

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 $f \in [1..m]$), since refining preceding attributes will certainly causes 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: Additional Experiments

C.1 Performance evaluation

Additional experiments reporting the execution time and the number of reported signficant pattern by DEvIANT on Movielens, Yelp, CHUS and EPD8. In this

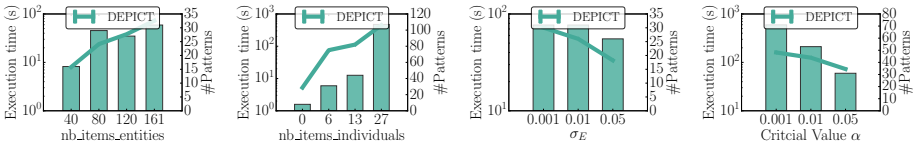


Fig. 5: Effectiveness of DEvIANT on Movielens when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$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

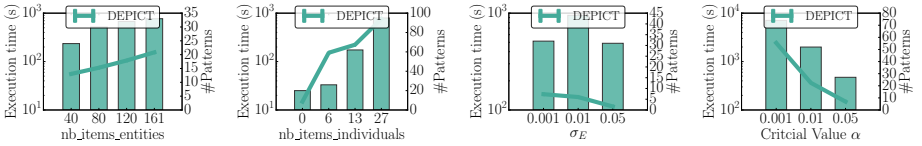


Fig. 6: Effectiveness of DEvIANT on Movielens when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$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

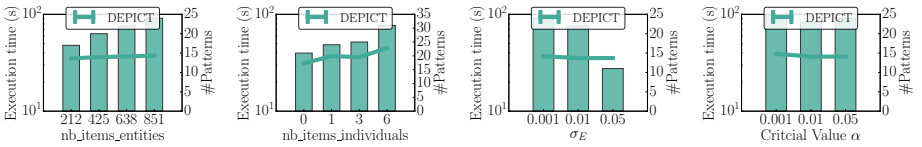


Fig. 7: Effectiveness of DEvIANT on Yelp when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha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

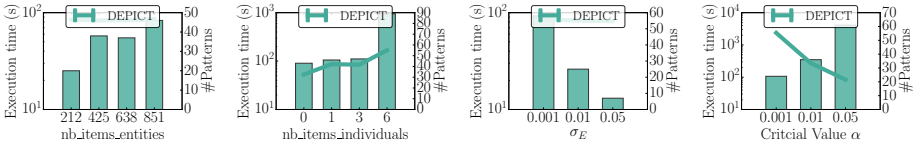


Fig. 8: Effectiveness of DEvIANT on Yelp when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha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

experiments we study also 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.

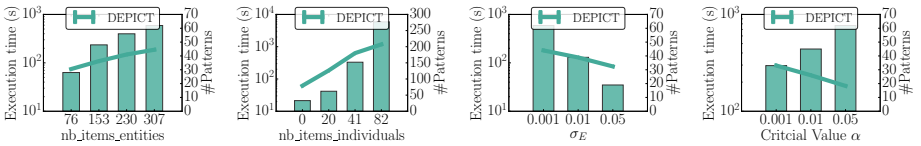


Fig. 9: Effectiveness of DEvIANT on CHUS when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled

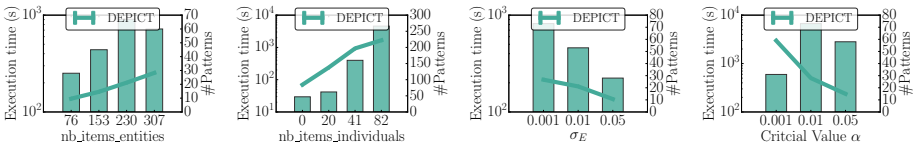


Fig. 10: Effectiveness of DEvIANT on CHUS when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is enabled

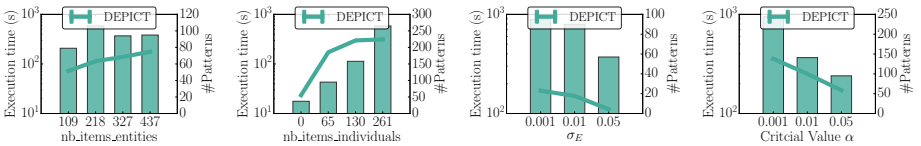


Fig. 11: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled.

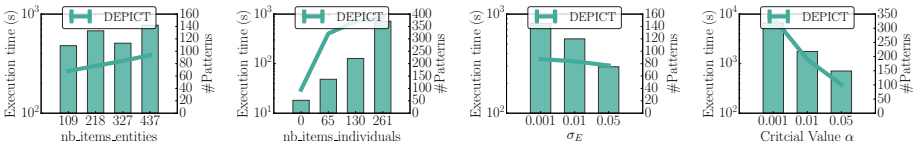


Fig. 12: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is enabled.

C.2 Qualitative evaluation

In this appendix, we report additional illustrative examples depicting the significant patterns discovered by DEvIANT when carried on the Eighth European Parliament (EPD8) dataset, US House of representatives (CHUS) dataset and Yelp dataset.

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 7: Top-5 Exceptional consensual/conflictual subjects among European Political Groups 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	*	10 Health & Medical	0.05	-0.33	< 0.001	Conflict
p_2	*	03 Automotive	0.05	-0.29	< 0.001	Conflict
p_3	*	14 Local Services	0.05	-0.25	< 0.001	Conflict
p_4	newcommer	21.177 American (Traditional) and 21.24 Breakfast & Brunch	-0.06	0.24	< 0.01	Consensus
p_5	*	11 Home Services	0.05	-0.25	< 0.01	Conflict

Table 8: Top-5 Exceptional consensual/conflictual subjects among Yelp users. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.