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 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) [28]/ Exceptional Model Mining (EMM) [6] 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 [21] 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 $X \geq \sigma$,

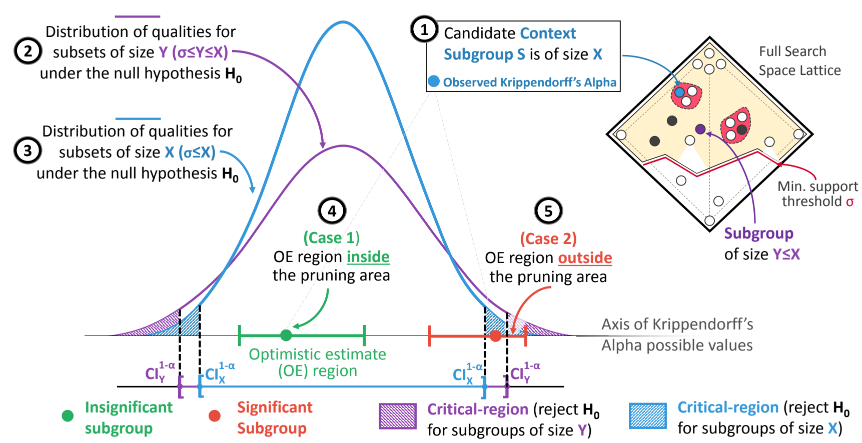


Fig. 1: Main DEvIANT properties for safe sub-search space pruning. A subgroup is reported as significant if its related Krippendorff's Alpha falls in the critical region of the corresponding empirical distribution of random subsets (DFD). When traversing the search space downward (decreasing support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

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. 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. [7,25]) 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-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) [15] 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 [21, 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 [21, p.232].

Discovering Significant Patterns. Statistical assessment of patterns has received attention for a decade [27,17], especially for association rules [16,26]. Some

work focused on statistical significance of results in SD/EMM during enumeration [7,25] or a posteriori [8] for statistical validation of the found subgroups.
Voting and Rating Data Analysis. Previous work [2] 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 [4] 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* $o : G_I \times G_E \to O$ *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 cat-

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

egorical, 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*, 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 [21]). 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 $c = \langle themes \supseteq \{7.30 \text{ Judicial Coop.}\}\rangle$, 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; $D_{\text{obs}}(e) = \sum_{o_1,o_2 \in O^2} \delta_{o_1 o_2} \dfrac{m_e^{o_1} \cdot m_e^{o_2}}{m_e \cdot (m_e - 1)}$

	[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 [6], while giving particular attention to the statistical significance and soundness of the discovered patterns [17].

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 [7,25,27], 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. [7]).

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. [7]). 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 [25].

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

[4] This paradigm naturally raises the question of how to address the *multiple comparisons problem* [19]. 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.

H_0. To easily manipulate θ_k, we reformulate A using Equations (1)-(3):

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

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 $CI_k^{1-\alpha}$ by computing $\mu_k = E[\theta_k]$ and $\sigma_k^2 = \mathrm{Var}[\theta_k]$. 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 $\mathrm{Var}[\theta_k]$. In the former case, the confidence interval $CI_k^{1-\alpha}$ endpoints are given by [14, 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 $\mathrm{Var}[\theta_k] = \binom{n}{k}^{-1} \sum_{S\in F_k} \left(\frac{\sum_{e\in S} v_e}{\sum_{e\in S} w_e} - E[\theta_k]\right)^2$.

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 [9] and [20, p.351], since no easy analytic expression of $E[\theta_k]$ and $\mathrm{Var}[\theta_k]$ 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:*

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

[5] In the same line of reasoning of [5], 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 $\mathcal{N}(\mu, \sigma^2)$ [3, p.413].

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

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

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; all appendices are available at https://hal.archives-ouvertes.fr/hal-02161309/document.

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 [15] on Krippendorff's alpha. Eppstein and Hirschberg [11] 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:*

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

[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.

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 $\mathrm{Var}[\theta_k]$ 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 $\widehat{CI}_k^{1-\alpha}$.

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

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

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 $|G_E^c| \geq |G_E^d| \geq \sigma_E$, we have:*

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

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 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 [18]: "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* [10] to empirically generate the sampling distribution of A and produce an empirical confidence interval $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$.

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 [18,21], 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 $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$ is defined by the percentiles of $\hat{B}$, i.e., $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}} = [\hat{A}_{\lfloor \frac{\alpha}{2} \cdot b \rfloor}, \hat{A}_{\lceil (1-\frac{\alpha}{2}) \cdot b \rceil}]$. We denote by $\mathrm{MCI}^{1-\alpha}$ (Merged CI) the confidence interval that takes into consideration both $CI^{1-\alpha} = [\mathrm{le}_1, \mathrm{re}_1]$ and $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}} = [\mathrm{le}_2, \mathrm{re}_2]$. We have $\mathrm{MCI}^{1-\alpha} = [\min(\mathrm{le}_1, \mathrm{le}_2), \max(\mathrm{re}_1, \mathrm{re}_2)]$.

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 [23]) and exhaustive (e.g., GP-growth [24]) enumeration algorithms exist. We exhaustively enumerate all candidate subgroups while leveraging closure operators [12] (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 [2], we apply EnumCC to enumerate subgroups g (resp. c) in $\mathcal{D}_I$ (resp. $\mathcal{D}_E$). EnumCC follows the line of algorithm CloseByOne [22]. 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 $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}}$ is calculated.

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 $OE(G_E^c, \sigma_E^g) \subseteq \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$, the search subspace under c can be pruned. Otherwise, $A^g(G_E^c)$ is computed and evaluated against $\mathrm{MCI}^{1-\alpha}_{|G_E^c|}$. If $A^g(G_E^c) \notin \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$, 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: $\forall p' = (g', c') \in \mathcal{P} \setminus P : \textit{p-value}^{g'}(c') \leq \alpha \Rightarrow \exists p = (g, c) \in P$ s.t. $\mathrm{ext}(q) \subseteq \mathrm{ext}(p)$, with $\mathrm{ext}\,(q = (g', c')) \subseteq \mathrm{ext}\,(p = (g, c))$ 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.

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

	$\|G_E\|$	$\mathcal{A}_E$ (Items-Scaling)	$\|G_I\|$	$\mathcal{A}_I$ (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%	$\simeq 1.14$

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 [13]).

$\mathbf{Q}_1$. 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 $\mathrm{ECI}_k^{1-\alpha} = [a^E, b^E]$. 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 $\mathrm{ECI}_k^{1-\alpha}$. 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.

$\mathbf{Q}_2$. 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

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

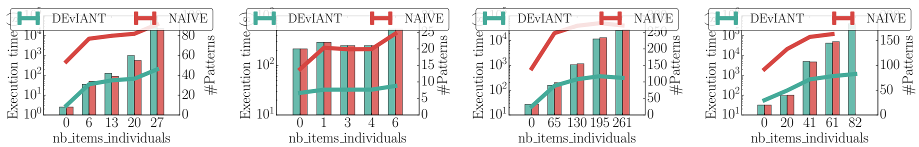
[9] 102^{nd}-115^{th} congresses of the US House of Representatives (Period: 1991-2015).

[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 $\alpha = 0.05$.

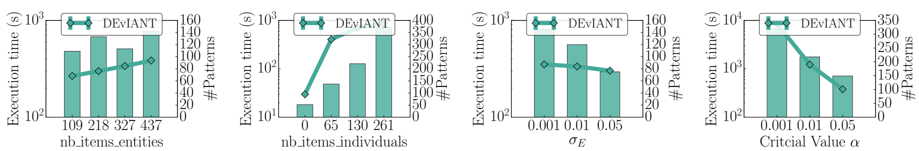


Fig. 3: Effectiveness of DEvIANT on EPD8 when varying sizes of both search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full search spaces $\mathcal{D}_E$ and $\mathcal{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha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

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 $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$ 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 $\mathcal{D}_I$ leads to more candidate groups of individuals g which require DEvIANT to: (i) generate $\mathrm{CI}^{1-\alpha}_{\mathrm{bootstrap}}$ 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

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

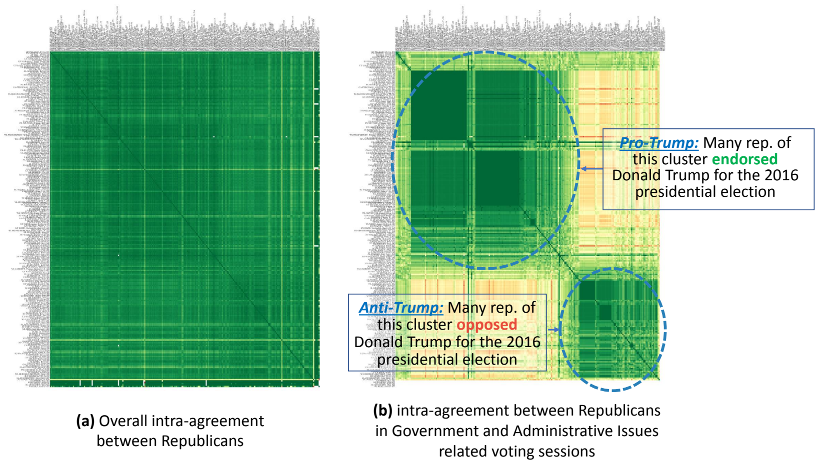


Fig. 4: Similarity matrix between Republicans, illustrating Pattern p_1 from Table 5. Each cell represents the ratio of voting sessions in which Republicans agreed. Green cells report strong agreement; red cells highlight strong disagreement.

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$^{\text{th}}$ 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

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

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, 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 [17], (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 [2].

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.

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. A. Belfodil, S. Cazalens, P. Lamarre, and M. Plantevit. Flash points: Discovering exceptional pairwise behaviors in vote or rating data. ECML/PKDD, 2017.
3. T. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

4. M. Das, S. Amer-Yahia, G. Das, and C. Yu. Mri: Meaningful interpretations of collaborative ratings. *PVLDB*, 4(11):1063–1074, 2011.
5. T. De Bie. An information theoretic framework for data mining. KDD, 2011.
6. W. Duivesteijn, A. J. Feelders, and A. Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.
7. W. Duivesteijn and A. Knobbe. Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. ICDM, 2011.
8. W. Duivesteijn, A. J. Knobbe, A. Feelders, and M. van Leeuwen. Subgroup discovery meets bayesian networks - an exceptional model mining approach. ICDM, 2010.
9. 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.
10. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
11. D. Eppstein and D. S. Hirschberg. Choosing subsets with maximum weighted average. *J. Algorithms*, 24(1):177–193, 1997.
12. B. Ganter and S. Kuznetsov. Pattern structures and their projections. ICCS,2001.
13. B. Ganter and R. Wille. *Formal concept analysis - mathematical foundations*. Springer, 1999.
14. S. Geisser. *Predictive Inference*, volume 55. CRC Press, 1993.
15. H. Grosskreutz, S. Rüping, and S. Wrobel. Tight optimistic estimates for fast subgroup discovery. ECML/PKDD, 2008.
16. W. Hämäläinen. Statapriori: an efficient algorithm for searching statistically significant association rules. *Knowl. Inf. Syst.*, 23(3):373–399, 2010.
17. W. Hämäläinen and G. I. Webb. A tutorial on statistically sound pattern discovery. *Data Min. Knowl. Discov.*, 33(2):325–377, 2019.
18. 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.
19. S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
20. M. Kendall, A. Stuart, and J. Ord. Kendall's advanced theory of statistics. v. 1: Distribution theory. 1994.
21. K. Krippendorff. Content analysis, an introduction to its methodology. 2004.
22. S. O. Kuznetsov. Learning of simple conceptual graphs from positive and negative examples. PKDD, 1999.
23. M. van Leeuwen and A. J. Knobbe. Diverse subgroup set discovery. *Data Min. Knowl. Discov.*, 25(2):208–242, 2012.
24. F. Lemmerich, M. Becker, and M. Atzmueller. Generic pattern trees for exhaustive exceptional model mining. ECML/PKDD, 2012.
25. F. Lemmerich, M. Becker, P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Mining subgroups with exceptional transition behavior. KDD, 2016.
26. 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.
27. G. I. Webb. Discovering significant patterns. *Machine learning*, 68(1):1–33, 2007.
28. S. Wrobel. An algorithm for multi-relational discovery of subgroups. PKDD, 1997.