Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: http://www.tandfonline.com/loi/nnmr20

Supervised descriptive pattern discovery in Native American music

Kerstin Neubarth, Daniel Shanahan & Darrell Conklin

To cite this article: Kerstin Neubarth, Daniel Shanahan & Darrell Conklin (2018) Supervised descriptive pattern discovery in Native American music, Journal of New Music Research, 47:1, 1-16, DOI: 10.1080/09298215.2017.1353637

To link to this article: https://doi.org/10.1080/09298215.2017.1353637

JOURNAL OF NEW MUSIC RESEARCH, 2018
VOL. 47, NO. 1, 1–16
https://doi.org/10.1080/09298215.2017.1353637

Routledge
Taylor & Francis Group

Supervised descriptive pattern discovery in Native American music

Kerstin Neubarth[a], Daniel Shanahan[b] and Darrell Conklin[c,d]

[a]Canterbury Christ Church University, UK; [b]Louisiana State University, Louisiana, USA; [c]University of the Basque Country UPV/EHU, Spain; [d]IKERBASQUE, Basque Foundation for Science, Spain

ABSTRACT

The discovery of recurrent patterns in groups of songs is an important first step in computational corpus analysis. In this paper, computational techniques of supervised descriptive pattern discovery are applied to model and extend ethnomusicological analyses of Native American music. Using a corpus of over 2000 songs collected and transcribed by anthropologist Frances Densmore and building on Densmore's own music content features, the analysis identifies musical differences between indigenous groups and between musical style areas of the North American continent. Contrast set mining is adapted to discover global-feature patterns which are distinctive for a group, statistically significant and maximally general. The work extends previous descriptive studies in computational folk music analysis by considering feature-set patterns of variable size. Discovered patterns confirm, differentiate and complement ethnomusicological observations on Native American music.

ARTICLE HISTORY
Received 12 February 2017
Accepted 27 June 2017

KEYWORDS
Contrast pattern mining; contrast set mining; emerging pattern mining; computational music analysis; corpus analysis; folk music analysis; Native American music

1. Introduction

Many collections of folk and indigenous music are accompanied by metadata which organise the collection into labelled groups of songs, such as tune families, folk music genres, geographical regions or language families (e.g Conklin & Anagnostopoulou, 2011; Conklin, 2013b; Shanahan & Shanahan, 2014; Volk & van Kranenburg, 2012). Not surprisingly, then, computational folk music analysis has explored *supervised* data mining techniques which take into account group labels, including attribute assessment and selection (Hillewaere, 2013; van Kranenburg, Volk, & Wiering, 2013), classification (Conklin, 2013b; Martins & Silla, 2015), subgroup discovery (Taminau et al., 2009) and emerging pattern mining (Conklin, 2010). Among these, *descriptive* mining methods, such as subgroup discovery or emerging pattern mining, aim to make patterns explicit which differentiate between groups. Supervised descriptive pattern discovery is often used for explorative analysis, suggesting patterns for further human interpretation (Novak, Lavrač, & Webb, 2010); descriptive patterns are generally evaluated by interestingness (Klösgen, 1996), and discovered patterns should be comprehensible (Novak, Lavrač, & Webb, 2009). The research presented in this paper applies supervised descriptive analysis to the Densmore collection of Native American music to identify musical features which capture contrasts between songs of different indigenous groups or tribes and different musical style areas in North American native music.

Native American music has long been the subject of ethnomusicological study, including quantitative and statistical analysis. Comparative studies have analysed repertoires by tribes (e.g. Densmore, 1923; Herzog, 1936), song types (e.g. Gundlach, 1932; Herzog, 1935b) or the age of songs (Densmore, 1918). In-depth studies have focused on stylistic differences between song types within tribes (e.g. Densmore, 1913; Nettl, 1955; Nettl, 1967; Nettl & Blum, 1968) or tribal and regional variation of song types and performance practices (e.g. Hatton, 1986). On the other hand, Nettl 1954 identified geographical areas of relatively homogeneous musical styles shared by several tribes (see also Levine, 1998; Vennum, 2000), and Herzog (1935b) postulated special song types which cut across tribal repertoires (see also Hatton, 1986). Systematic catalogues of analysis criteria or features support both qualitative comparisons (e.g. Herzog, 1936; Nettl, 1954) and quantitative analyses (e.g. Densmore, 1910; Densmore, 1957), generally covering aspects of melodic analysis (e.g. tone material, melodic motion), rhythmic-metric analysis (e.g. dominant tone durations, metrical organisation), musical form (e.g. phrase and repetition structures) and performance style (e.g. singing style, instrumental accompaniment). Quantitative analyses are based on tabulations of features against individual

CONTACT Darrell Conklin darrell.conklin@ehu.eus Darrell Conklin, University of the Basque Country UPV/EHU, San Sebastián, Spain

songs, absolute or relative feature frequencies for groups of songs, descriptive statistics or rank order analysis (e.g. Gundlach, 1932; Herzog, 1935a, 1935b).

Frances Densmore collected, transcribed and published over 2000 songs of various North American tribes, accompanied by detailed contextual information and, in many writings, quantitative melodic and rhythmic analyses. To illustrate Densmore's analysis by an example, Figure 1 reproduces two tables from Densmore's study of Northern Ute music (Densmore, 1922), which focus on the melodic structure of the transcribed songs: the relation between accented melodic tones. The first table lists against each of the possible attribute values the songs exhibiting the relevant structure, for a subgroup of 16 Ute war songs; the second table presents the frequencies of attribute values for the analysed Ute songs compared against the combined frequencies for previously published songs of the Chippewa and Teton Sioux. In her comments, Densmore observes that the 'percentage of songs in mixed form is more than twice as great in the Ute than in the Chippewa and Sioux' (Densmore, 1922, p. 53). This analysis effectively constitutes an example of supervised descriptive pattern discovery: a corpus organised into groups, extraction of features for the songs in the corpus, determination of feature frequencies, and comparison of a target group against the combined remaining groups known as one-vs-all comparison, in this particular case quantifying the difference by the percentage ratio—'more than twice as great'—corresponding to growth rate in emerging pattern mining (Conklin, 2010; Dong & Li, 1999).

In this paper we build on Densmore's seminal work: applying computational techniques to the digitised songs (Section 2.1), we identify significant associations between Densmore's music content descriptors and tribal repertoires, but also consider an alternative grouping into musical areas (Section 2.2) and additional computationally extracted features (Section 2.3). Following from Densmore's analyses, the analysis applies song-level, global features (attribute—value pairs) and draws on emerging pattern and contrast set mining to identify and evaluate potentially interesting feature sets (Section 3). Extracted patterns are discussed in the context of existing ethnomusicological work (Section 4). Our case study of applying supervised descriptive pattern discovery to analyse a large collection of Native American music demonstrates the potential of descriptive data mining methods in comparative and corpus-level music analysis.

2. Corpus and representation

This section introduces the Densmore corpus analysed in the current study, summarising relevant metadata and outlining the music content features used. Particular attention is given to issues of data quality such as relevance of the data-set, contextual information and missing values.

2.1. The Densmore collection of Native American music

Frances Densmore studied Native American music over seven decades, from 1893 to the late 1950s, recording, transcribing and publishing songs of Native American tribes across the continent. Most of this work was carried out under the auspices of the Smithsonian Institution's Bureau of American Ethnology (Smithsonian Institution, 1971); the volumes on Cheyenne and Arapaho and on Maidu music were supported by the Southwest Museum (Densmore, 1936; Densmore, 1958). While collected songs attempt to give a representative picture of a tribe's songs (e.g. Densmore, 1932a, p. 19; Densmore, 1939, p. 43), Densmore was particularly interested in old songs remembered by older members of the tribe (Densmore, 1915, p. 189). During the collection of Mandan and Hidatsa songs, 'singers were encouraged to suggest the songs which they regarded as valuable for preservation' (Densmore, 1923, p. 12). As a record of Native American music, the material is particularly remarkable in that Densmore 'heard and recorded songs, and attended ceremonies, which no white person ever before had been allowed to hear or see' (Hofmann, 1946, p. 47). Densmore's transcriptions and recordings have been used in both contemporary and more recent studies of Native American music and dance (e.g. Browner, 2000; Gundlach, 1932; Herzog, 1935b; Kurath, 1953; Nettl, 1954).

The analyses presented in this paper are based on Densmore's published transcriptions of over 2000 songs. Most of them have been encoded in Humdrum (Shanahan & Shanahan, 2014). In order to extract computational features using existing tools, all Humdrum files were further converted into MIDI. The digitised collection covers 16 books, generally encoding all transcriptions provided in a book; for the Mandan and Hidatsa and the Nootka and Quileute books the digitisation is incomplete. Humdrum and consequently MIDI encodings are available for 2083 out of 2218 songs.

In Densmore's publications, the transcriptions are accompanied by extensive introductions to the context of songs and, in the majority of the books, musical analyses: for sets of melodic and rhythmic music content attributes, Densmore manually extracted attribute values for each song, which form the basis of comparative quantitative analyses (see example in Figure 1). These writings inform the organisation of the corpus into groups, allow the use of Densmore's own features for the description of songs

WAR SONGS

67. War song (a)
68. War song (b)
69. War song (c)
70. War song (d)
71. Scout song
... ...
82. War song (i)

STRUCTURE

	Number of songs	Serial Nos. of songs
Melodic	8	68, 70, 71, 72, 74, 77, 81, 82.
Melodic with harmonic framework	6	67, 69, 73, 76, 78, 80.
Harmonic	2	75, 79.
Total	16	

STRUCTURE

	Ute songs		Chippewa and Sioux songs		Chippewa, Sioux, and Ute songs	
	Number	Per cent	Number	Per cent	Number	Per cent
Melodic [a]	54	*49*	397	*66*	451	*64*
Melodic with harmonic framework [b]	32	*29*	85	*14*	117	*16*
Harmonic [c]	24	*22*	116	*19*	140	*20*
Irregular	....	....	2	....	2	....
Total	110	....	600	....	710	....

[a] Songs are thus classified if contiguous accented tones do not bear a simple chord-relation to each other.
[b] Songs are thus classified if only a portion of the contiguous accented tones bear a chord-relation to each other.
[c] Songs are thus classified if contiguous accented tones bear a simple chord-relation to each other.

Figure 1. Tabular analysis of Ute songs by Frances Densmore: analysis of melodic structure. Top left: index of Ute war songs, selection (Densmore, 1922, pp. 12–13). Top right: analysis of Ute war songs by serial number (Densmore 1922, p. 162). Bottom: comparison of Ute songs against previously analysed songs (Densmore, 1922, p. 38).

in addition to computationally extracted features, and provide references both for the assessment of the data-set and the interpretation of results.

2.2. Organisation of the data-set

Based on Densmore's publications of transcribed songs, the corpus is mainly organised by Native American tribes. Within the published volumes, songs are often further grouped according to their use or song type; alternative organisations include grouping by reservations (Densmore, 1910), tribal societies (Densmore, 1923) or, in case of some volumes covering more than one tribe, the individual tribes (Densmore, 1957). In this study, the focus of analysis lies on comparing repertoires by tribe groups and by musical areas which capture stylistic characteristics shared by several tribes.

2.2.1. Grouping by tribes

Densmore generally dedicated individual publications of transcribed songs to the repertoire of one tribe. A few books cover more than one repertoire. For example, the 1923 Bulletin presents Mandan songs together with some songs of the Hidatsa, including 'many songs which the tribes appear to have in common' (Densmore, 1923, p. 12). In Densmore's comparative analyses Mandan and Hidatsa are treated as one group, as are Yuman and Yaqui (Densmore, 1932b), Nootka and Quileute (Densmore, 1939), and the Pueblo tribes Acoma, Isleta, Chochiti and Zuñi (Densmore, 1957). For the songs collected in British Columbia, Densmore recorded the singers' home localities but did not further specify their tribes (Densmore, 1943b).

The repertoire indicated in the book title forms the focus of analysis, but—from the Northern Ute book

onwards—is also compared against previously analysed songs, e.g. Ute songs against Chippewa and Sioux songs (Densmore, 1922; see also Figure 1); Papago songs against Chippewa, Sioux, Ute, Mandan and Hidatsa songs (Densmore, 1929a); or Menominee songs against Chippewa, Sioux, Ute, Mandan and Hidatsa, Papago and Pawnee songs (Densmore, 1932a). In this paper, we follow Densmore's publications in order to group the corpus by tribes—merging the Chippewa I and II books into one group of Chippewa songs—but replace Densmore's chronologically cumulative comparison by considering all 15 tribe groups in the analysis.

2.2.2. Grouping by musical areas

In 1954, Nettl presented an extensive survey of North American native music based on a wealth of published and unpublished material (transcriptions, field recordings, descriptions), including publications by Densmore. The survey identified *musical areas* in Native American music: 'A musical area may be defined as a geographic area whose inhabitants share in a generally homogeneous musical style. Such an area is unified by one or several important traits which are not found with the same degree of intensity in neighboring areas' (Nettl, 1954, p. 46). The characterisation of styles is 'defined by statements of frequency—statistical statements—rather than by statements which indicate only the absolute presence or absence of a given trait' (p. 47), with an 'emphasis on the technical aspects of musical style' (p. 46), such as melodic style (range, scale, melodic movement), rhythm (durational values used, metre, instrumental accompaniment), form (overall structure, number and lengths of sections, form types) and other features (vocal technique, dynamics, accentuation, instruments). Of the six musical areas proposed by Nettl, five are represented in the Densmore corpus of the current study. Figure 2 gives an overview, mapping tribes onto areas.

2.3. Music content representation

Following Densmore's own analyses, the current study focuses on patterns which comprise *global features*: features that summarise songs by attribute—value pairs. Features are derived from Densmore's tabular analyses and through computational feature extraction from the encoded songs.

2.3.1. Densmore features

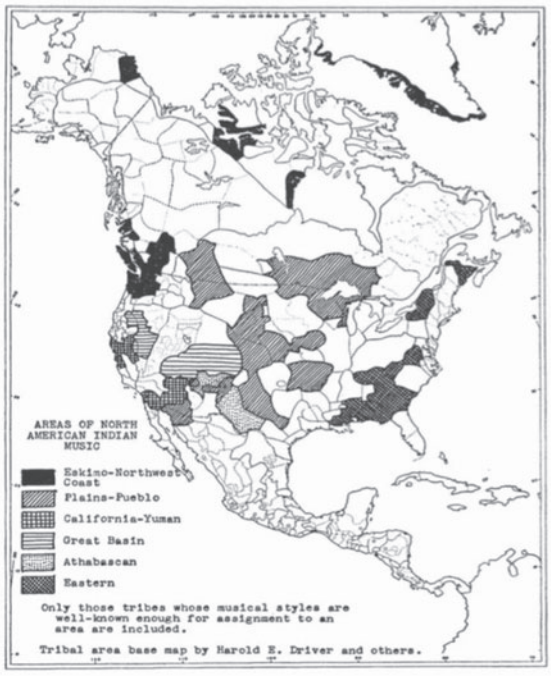
Of the 16 Densmore books supporting the corpus of this study, 13 books include tabular analyses of the transcribed songs (see example in Figure 1), with the exception of the Cheyenne and Arapaho, British Columbia and Seminole books. These tables were manually translated into content descriptions for the analysed songs, assigning to each song all features satisfied by the song.

The number of attributes analysed by Densmore ranges from five (Maidu) to 18 (Teton Sioux and Northern Ute): Table 1 summarises the coverage of attributes across the 13 books containing tabular analyses. Checkmarks highlight the attributes included in Densmore's tables. Attributes marked by a diamond were manually added by the authors, generally based on analysis of the published scores. For Chippewa I, the number of rhythmic units (attribute rhythmUnit) was extracted from Densmore's listing of identified units (Densmore, 1913, pp. 309–332). No attempt was made to complete missing values for attributes which involve some interpretation on Densmore's part, such as those requiring the determination of a keynote. The attribute key and the attributes relating to tempo of voice, rhythmic accompaniment and their relationship were excluded from the computational analysis as they are unavailable for the majority of songs in the corpus.

Densmore's partitioning of attributes into values varies between the different books. For example, in analysing the relation between the last tone of a song and the range of the song (attribute lastReCompass), earlier books like those on Teton Sioux, Northern Ute or Pawnee distinguish e.g. 'songs containing a fourth below the final tone' and 'songs containing a major third below the final tone', while the Nootka and Quileute and the Choctaw books integrate those constellations into a more general value 'songs containing lower tones than the final tone'. Attribute values were thus harmonised across the data-set. The resulting features are summarised in Table 2.

2.3.2. Computational features

The Densmore features are complemented with computational features extracted from the MIDI encodings of songs. For the purposes of this study, we selected 10 jSymbolic attributes (McKay, 2010) which reflect analysis criteria of existing ethnomusicological studies (e.g. Herzog, 1928; Gundlach, 1932; Herzog, 1936; Nettl, 1954), such as average size of melodic intervals (AverageMelodicInterval); the frequency of certain melodic interval classes (RepeatedNotes, StepwiseMotion, ChromaticMotion, MelodicThirds and MelodicTritones); and aspects of melodic contour like proportion of ascending intervals (DirectionofMotion) or duration and ambitus of melodic arcs (MelArcDur and MelArcSize). Tempo is not explicitly represented in the corpus: only a fraction of Humdrum files are annotated with tempo. When converting into MIDI, all songs were standardised to ♩ = 60. With MIDI files exported from symbolic scores, average note duration (AverageNoteDuration) then captures aspects of the rhythmic density of songs.



Eskimo–Northwest Coast Area (309)

British Columbia (98), Nootka and Quileute (211).

Plains–Pueblo Area (1292)

Chippewa (383), Menominee (144), Pawnee (86), Teton Sioux (245), Mandan and Hidatsa (111), Cheyenne and Arapaho (72), Pueblo tribes (82), Papago (169).

California–Yuman Area (185)

Yuman and Yaqui (132), Maidu (53).

Great Basin Area (116)

Northern Ute (116).

Eastern Area (316)

Choctaw (69), Seminole (247).

Figure 2. Musical areas in North American native music. Left: map of musical areas (from Nettl (1954, p. 45), reprinted with permission of the American Folklore Society). Right: areas and tribes represented in the Densmore corpus (counts in brackets indicate the number of songs in each area and tribe).

Table 1. Attributes analysed by Densmore, for melodic analysis (top) and rhythmic analysis (bottom). ✓: attribute covered by Densmore's tabular analysis. ◇: analysis added by the authors. The final column shows the number of songs in the corpus that have defined values for each attribute, for the attributes that have been encoded. — : attribute not included in the computational analysis. The attributes used in this study are further described in Table 2.

Attribute	Chippewa I	Chippewa II	Teton Sioux	Northern Ute	Mandan and Hidatsa	Papago	Pawnee	Menominee	Yuman and Yaqui	Nootka and Quileute	Choctaw	Pueblos	Maidu	Songs with attribute
tonality	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1738
firstReKey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		1686
lastReKey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		1686
lastReCompass	◇	◇	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	◇	1752
compass			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1392
material	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1723
accidentals	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				1539
structure	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				1539
firstDir	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	◇	1738
key			✓	✓										—
firstMetr	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	◇	1694
initMetre	◇	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1697
metreChange	◇	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1697
rhythmUnit	◇	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		1703
rhythmAccomp			✓	✓	✓		✓		✓		✓			—
tempoVoice			✓	✓										—
tempoAccomp			✓	✓										—
tempoRelation	✓	✓	✓	✓										—

The selected jSymbolic attributes are numeric attributes. To derive categorical features, extracted features are discretised into two bins, Low and High, with a split point at the mean value in the digitised corpus (cf. Taminau et al., 2009).

For jSymbolic attributes, only those songs not having a Humdrum entry in the corpus have missing values. Thus 2083 songs contain values for all considered jSymbolic attributes. For convenience, to distinguish Densmore and jSymbolic attributes, the former will be written with small initial letter (e.g. tonality) and the latter with capital initial letter (e.g. DirectionofMotion).

Example: To illustrate the data and, subsequently, data mining methods, we introduce a small example data-set

Table 2. Densmore features included in the computational analysis.

Attribute	Description	Values
tonality	tonality: third above keynote	major, minor, both, third lacking, irregular
firstReKey	first note relative to keynote	keynote, triad tone within octave, other tone within octave, octave, triad tone above octave, other tone above octave, irregular
lastReKey	last note relative to keynote	keynote, fifth, third, second, sixth, irregular
lastReCompass	last note relative to compass of song	lowest, containing lower (but not the highest tone), highest, irregular
compass	number of tones comprising compass	4 or less, 5 to 8, 9 to 12, more than 12
material	tone material	second 5-toned scale, fourth 5-toned scale, octave complete, other combinations
accidentals	accidentals: chromatic alterations of tones	triad tone(s) altered, other tone(s) altered, triad and other tone altered, none or unidentified
structure	relation between contiguous accented tones	melodic (incl. irregular), harmonic, mixed
firstDir	direction of first progression	downward, upward
firstMetr	part of measure on which song begins	accented, unaccented, without accents
initMetre	rhythm (metre) of first measure	duple, triple, quintuple, other, irregular
metreChange	change of time (measure-lengths)	yes, no, without accents
rhythmUnit	rhythmic unit(s) of song	yes, no

Table 3. A small example data-set, using some of the attributes presented in Section 2. Each song has a song identifier, one group attribute (tribe) and four content attributes. Missing values are indicated by –.

ID	tribe	tonality	structure	initMetre	DirectionofMotion
124	chippewa	minor	melodic	triple	Low
126	chippewa	minor	melodic	–	Low
338	chippewa	major	mixed	triple	Low
811	mandan	major	mixed	duple	Low
812	mandan	minor	harmonic	triple	Low
830	mandan	minor	mixed	triple	–
1333	nootka	minor	melodic	duple	–
1334	nootka	major	melodic	triple	High
2201	choctaw	major	–	duple	High
2202	choctaw	major	–	duple	Low

(Table 3): a subset of 10 songs from the Densmore corpus, showing a selection of attributes. Each song is identified by its catalogue number in the Densmore collection (ID). A group attribute (tribe) organises the data-set into four tribe groups: values chippewa, mandan, nootka and choctaw. Additionally, each song is described by four content attributes: three Densmore attributes (tonality, structure and initMetre) derived from Densmore's analyses and one jSymbolic attribute (DirectionofMotion) extracted from the MIDI files. One of the Mandan songs and one of the Nootka songs have not been digitally encoded and thus do not have a value for DirectionofMotion. The Choctaw songs miss values for attribute structure, as this attribute is not covered by Densmore's tabular analyses of Choctaw songs (see Table 1). For the Chippewa song ID : 126 Densmore only transcribed the melodic outline and did not apply rhythmic-metric attributes, hence the song does not have a value for the attribute initMetre.

3. Supervised descriptive pattern discovery

Supervised descriptive pattern discovery is the task of learning patterns of interest from labelled data (Novak, Lavrač, & Webb, 2009). Applied to music, it is an example of *inter-opus*, rather than *intra-opus*, pattern discovery (Conklin, 2010): the aim is to find interesting patterns which reoccur across several pieces in a corpus. More specifically, in this study we are interested in global-feature patterns which capture distinctive characteristics of groups in the Densmore collection.

3.1. Patterns

A *pattern* is a set of content features. In this paper we consider *global-feature patterns*, i.e. sets of global features. A *global feature* is an attribute—value pair $a : v$ consisting of an attribute a and value v. Specific attributes and features will be indicated in sans serif font, e.g. the attribute tonality or the feature tonality : major. A piece *satisfies* a global feature $a : v$ if it has value v for attribute a. A piece satisfies a pattern if it satisfies all features in the set. The number of pieces satisfying a pattern gives the *support count* of the pattern.

A pair $X \rightarrow G$ comprising a pattern X and a single group feature G is called an *association*. The *relative support* of a pattern in a group corresponds to its empirical

Table 4. Contingency table for an association $X \rightarrow G$.

	G	$\neg G$	sum
X	$n(X, G)$	$n(X, \neg G) = n(X) - n(X, G)$	$n(X)$
$\neg X$	$n(\neg X, G) = n(G) - n(X, G)$	$n(\neg X, \neg G) = n(\neg G) - n(X, \neg G)$	$n(\neg X)$
sum	$n(G)$	$n(\neg G)$	N

conditional probability $\mathbb{P}(X \mid G) = n(X, G)/n(G)$. The occurrence of a pattern X with respect to a given group G and the background $\neg G$ of all groups other than G can be summarised in a 2×2 contingency table (Table 4). The marginal counts $n(X)$ and $n(G)$ refer to the support counts of pattern X and group G in the data-set, i.e. the number of pieces satisfying pattern X and the number of pieces in group G; N denotes the total number of pieces in the data-set. The inner cells quantify the frequency of pattern X in the target group, $n(X, G)$, i.e. the number of pieces in group G which satisfy pattern X, and in the background, $n(X, \neg G)$, i.e. the number of pieces outside group G which satisfy pattern X. When $N, n(X), n(G)$ and $n(X, G)$ are known, all other counts in the contingency table can be derived.

To illustrate these concepts, consider the example data-set of Table 3. The global feature tonality : major, and consequently the singleton feature set {tonality : major}, are satisfied by five songs in the data-set, and the feature set {tonality : major , initMetre : duple} is satisfied by three songs (Table 5). The distribution of the pattern {tonality : major , initMetre : duple} with respect to the group tribe : choctaw is summarised in the contingency table shown in Table 6, which can be filled after determining the support count of the group $n(G) = 2$, the pattern $n(X) = 3$, the association $n(X, G) = 2$, and the size of the data-set $N = 10$.

3.2. Identifying interesting patterns

Descriptive discovery methods often find large numbers of patterns. Pattern interestingness measures are thus applied to evaluate candidate patterns, reduce the search space during the mining process, and select or rank patterns in a way that identifies potentially interesting patterns for further inspection (Geng & Hamilton, 2006; Klösgen, 1996; Nguyen et al., 2016; Webb, 2007). In this paper, a pattern X is considered *interesting* for a group G if it meets two conditions (e.g. Bay & Pazzani, 2001). First, the pattern should be *distinctive* for group G: its *growth rate* should be larger than a specified threshold θ. The growth rate, defined as $\mathbb{P}(X \mid G)/\mathbb{P}(X \mid \neg G)$, compares the relative support of a pattern in the target group against its relative support in the background (Dong & Li, 1999). Hence a pattern is more distinctive for a group if it has higher support in the group compared to the background, taking into account the size of the group relative to the background.

Second, the over-representation of a pattern X in a group G should be *statistically significant*, i.e. a statistical test should reject the null hypothesis that the pattern distribution across data instances is independent of the group membership of instances (Bay & Pazzani, 2001; Conklin, 2013a; Webb, 2007; Wu et al., 2016). More specifically, a pattern is significant if its p-value, computed with Fisher's exact test (see Appendix 1), is lower than a specified significance level α. The smaller the p-value calculated by the test, the more surprising is the observed frequency of the pattern in the target group given the marginal counts in the contingency table. While growth rate quantifies in an intuitive way the magnitude of the difference in pattern support between a group and the background, Fisher's exact test ensures that the pattern is likely to capture a true, or statistically meaningful, difference between the group and the background (Bay & Pazzani, 2001; Webb, Butler, & Newlands, 2003).

3.3. Controlling the false discovery rate

In exploratory data analysis of a corpus with the size and dimensionality of the Densmore corpus, many false discoveries can be found (Webb, 2007; Liu, Zhang, & Wong, 2011) and failure to control them can lead to the reporting of an overwhelming number of trivial and spurious patterns. In this work, two methods are used to control the false discovery rate (FDR).

A first source of false discoveries is the *multiple testing problem*, where n tested patterns meeting a significance level of α can be expected to contain $n \times \alpha$ false discoveries. To handle this effect, one can apply the Bonferroni adjustment to the specified α: if n patterns are tested for significance, α is adjusted to α/n. Although it is known that the Bonferroni adjustment can be conservative, in this study we are interested mainly in finding small numbers of highly significant patterns and inspecting them for musicological meaning.

To determine the effectiveness of the Bonferroni correction, a permutation method (Liu, Zhang, & Wong, 2011) was used to estimate the FDR, under the assumption that most patterns found in randomised data would be artefactual. Tribe labels were randomly redistributed over songs containing all content features (Section 2.3),

Table 5. Selected feature sets for the example data-set.

Feature sets	IDs	Support
{tribe : choctaw}	2201, 2202	2
{tonality : major}	338, 811, 1334, 2201, 2202	5
{initMetre : duple}	811, 1333, 2201, 2202	4
{tonality : major , initMetre : duple}	811, 2201, 2202	3
{tribe : choctaw , tonality : major , initMetre : duple}	2201, 2202	2

Table 6. Contingency table, with respect to the example data-set, for the association {tonality : major , initMetre : duple} → tribe : choctaw.

	tribe : choctaw	¬ tribe : choctaw	sum
{tonality : major , initMetre : duple}	$n(X, G) = 2$	1	$n(X) = 3$
¬ {tonality : major , initMetre : duple}	0	7	7
sum	$n(G) = 2$	8	$N = 10$

while maintaining the overall tribe counts, then all significant (with Bonferroni correction) patterns containing one feature were discovered and counted. This permutation procedure was repeated 1000 times and the number of patterns—all assumed to be false discoveries—was counted. For $\alpha = 0.1$ this estimated a mean FDR of 0.052; for $\alpha = 0.05$, 0.028; and for $\alpha = 0.01$, 0.006. These results suggest that while on the conservative side, the Bonferroni correction leads as desired to a FDR smaller than the specified α. In particular, the simulation indicates that at $\alpha = 0.01$—the level used for the results presented in Section 4—there are few false positives to be expected.

In addition, false discoveries can arise due to *missing values* for attributes in some songs. Songs with missing values may or may not satisfy X, so counting them can distort growth rates and lead to inaccurate statements about statistical significance. Thus, in an association $X \rightarrow G$ any songs in group G that are missing an attribute occurring in X should not be considered part of the group count $n(G)$, and the overall count N should include only those songs not missing any attribute occurring in X. Similarly, if songs in $\neg G$ are missing an attribute occurring in X, counts in the background need to be adjusted. Failing to consider missing values can lead to either an over-estimation or an under-estimation of the growth rate and statistical significance.

3.4. Filtering redundant patterns

To discover patterns in a corpus, standard algorithmic methods from the field of association mining (Agrawal, Imielinski, & Swami, 1993; see also Tan, Steinbach and Kumar, 2006) are employed. For every group in turn, the search for patterns proceeds level-wise from the empty pattern and adds features at every search step. At every step the growth rate and significance are computed, possibly adjusted to account for missing values.

As there can be many redundant distinctive and significant patterns in a corpus, to control the search space and reduce redundancy among discovered patterns, an additional requirement is imposed for patterns to also be *maximally general*. Maximally general patterns are those which are interesting and which are not subsumed by another interesting pattern (Conklin, 2010). For example, the pattern {tonality : major , initMetre : duple} is subsumed by the single-feature patterns {tonality : major} and {initMetre : duple}: all pieces satisfying the first set also satisfy the latter two sets. Referring to Figure 3, if {tonality : major} or {initMetre : duple} is interesting for group tribe : choctaw, the feature set {tonality : major , initMetre : duple} cannot be maximally general (even though it might be interesting); hence the set does not need to be tested and the search branch can be pruned. In addition to reducing the search space, mining for maximally general interesting global-feature patterns favours concise patterns (Atzmüller, 2015; Dong & Li, 1999; Loekito & Bailey, 2006).

4. Patterns in the Densmore collection

The described pattern discovery method was applied to the Densmore corpus of 2218 Native American songs, represented by global features derived from 13 Densmore and 10 jSymbolic attributes. Songs were grouped into 15 tribes and 5 musical style areas. Results are presented supported by the visualisation method proposed in Novak, Lavrač, & Webb (2009): bar charts (Figures 4 and 5) summarise the distribution of a pattern in the data-set, showing the relative support of a pattern in the target group (dark fill) and background (light fill). The visualisation complements the presentation of growth rate and p-value by not only graphically illustrating the support

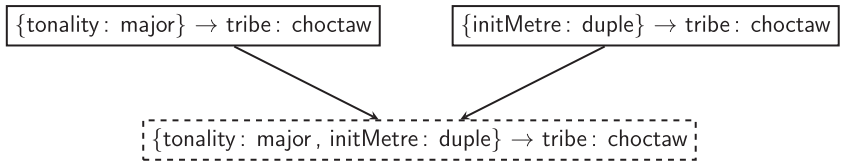


Figure 3. Subsumption relations among three associations. If either of the boxed associations is distinctive and significant, then the lower subsumed association cannot be a maximally general interesting pattern.

ratio but also indicating the frequency of the pattern in the data-set and in the target group and background, respectively.

4.1. Discovered single-feature patterns: Tribes

For the discussion of interesting patterns for tribes, Densmore's own analyses provide a useful reference. In order to relate discovered patterns to Densmore's observations, this section focuses on single-feature patterns and presents tribe descriptions in the chronological order of Densmore's cumulative analyses.

Figure 4 displays discovered interesting (distinctive and statistically significant) patterns which reflect features also pointed out by Densmore, e.g. 'In this *peculiarity* [proportion of songs in which the final tone is the lowest tone] the Mandan and Hidatsa songs show a larger proportion than any previously analyzed' (Densmore, 1923, p. 27, our emphasis; cf. lastReCompass : lowest → tribe : mandan). In Densmore's narrative, the comparison is sometimes quantified, e.g. 'the Papago contains [...] a *much larger percentage* on the fourth five-toned scale' (Densmore, 1929a, p. 12, our emphasis; cf. material : fourth_5-toned_scale → tribe : papago), 'the *highest per cent* [sic] ending on the keynote being found among the Chippewa' (Densmore, 1932a, p. 21, our emphasis; cf. lastReKey : keynote → tribe : chippewa) or even, corresponding to growth rate, 'The percentage of [Mandan and Hidatsa] songs [with mixed structure] is *four times* that of the Chippewa and *double* that of the Sioux' (Densmore, 1923, pp. 28–29, our emphasis; cf. structure : mixed → tribe : mandan). More often the text simply quotes the relative frequencies of the feature in the target group and in the background, without measuring the difference in frequencies, e.g. 'A change of measure lengths occurs in 97 per cent of these Pueblo songs and in only 80 per cent of the combined group' (Densmore, 1957, p. 112; cf. metreChange : yes → tribe : pueblos). Alternatively the difference is paraphrased without reference to the relative frequencies, e.g. 'The downward trend which characterised the songs previously analyzed is *less prominent* in the Nootka and Quileute songs' (Densmore, 1939, p. 43, our emphasis; cf. firstDir : up → tribe : nootka). Computational pattern discovery facilitates explicitly quantifying differences between groups in a consistent way.

The patterns presented in Figure 4 generalise Densmore's observations derived from a subset of the songs—according to the cumulative analyses—to the complete data-set, i.e. they point out interesting features of tribal repertoires that still hold if the target tribe is compared against additional tribes not included in Densmore's comparison. On the other hand, Table 7 gives examples of patterns highlighted by Densmore which on the full corpus are distinctive but not statistically significant (top) or neither distinctive nor significant (middle). For comparison, evaluation metrics for the partial corpus are also given. Only one of the listed patterns, firstDir : up for tribe : yuman, is distinctive and significant ($\alpha = 0.01$) on the smaller data-set. Several of the features suggested by Densmore are distinctive on both the smaller and complete data-sets, but not significant (at $\alpha = 0.01$ or even $\alpha = 0.1$). Generally, including tribes not considered by Densmore in her cumulative analysis does not drastically change results.

The bottom part of Table 7 lists some discovered patterns which complement Densmore's analyses either by suggesting an alternative feature value or by also including jSymbolic features in the analysis. In her analysis of Nootka and Quileute songs Densmore points out the percentage of songs ending on the third, 'exceeded only by the Yuman and Yaqui' (Densmore, 1939, p. 42), which is indeed distinctive, although not significant, also in our results. The computational analysis additionally suggests the ending on the second as a significantly distinctive feature for this group of songs (lastReKey : second → tribe : nootka). The pattern, however, is not frequent: only nine songs in the complete corpus end on the second, of which eight are Nootka and Quileute songs, a higher number than would be expected from the proportion of Nootka and Quileute songs in the corpus.

Although generally Densmore sought to collect representative songs of tribal repertoires, the Choctaw sample is dominated by dances (88% of Choctaw songs are dances), and certain features of Choctaw songs, such as the beginning on an accented part of the measure (firstMetr : accented, see Figure 4), may potentially be related to musical characteristics of dances (Densmore,

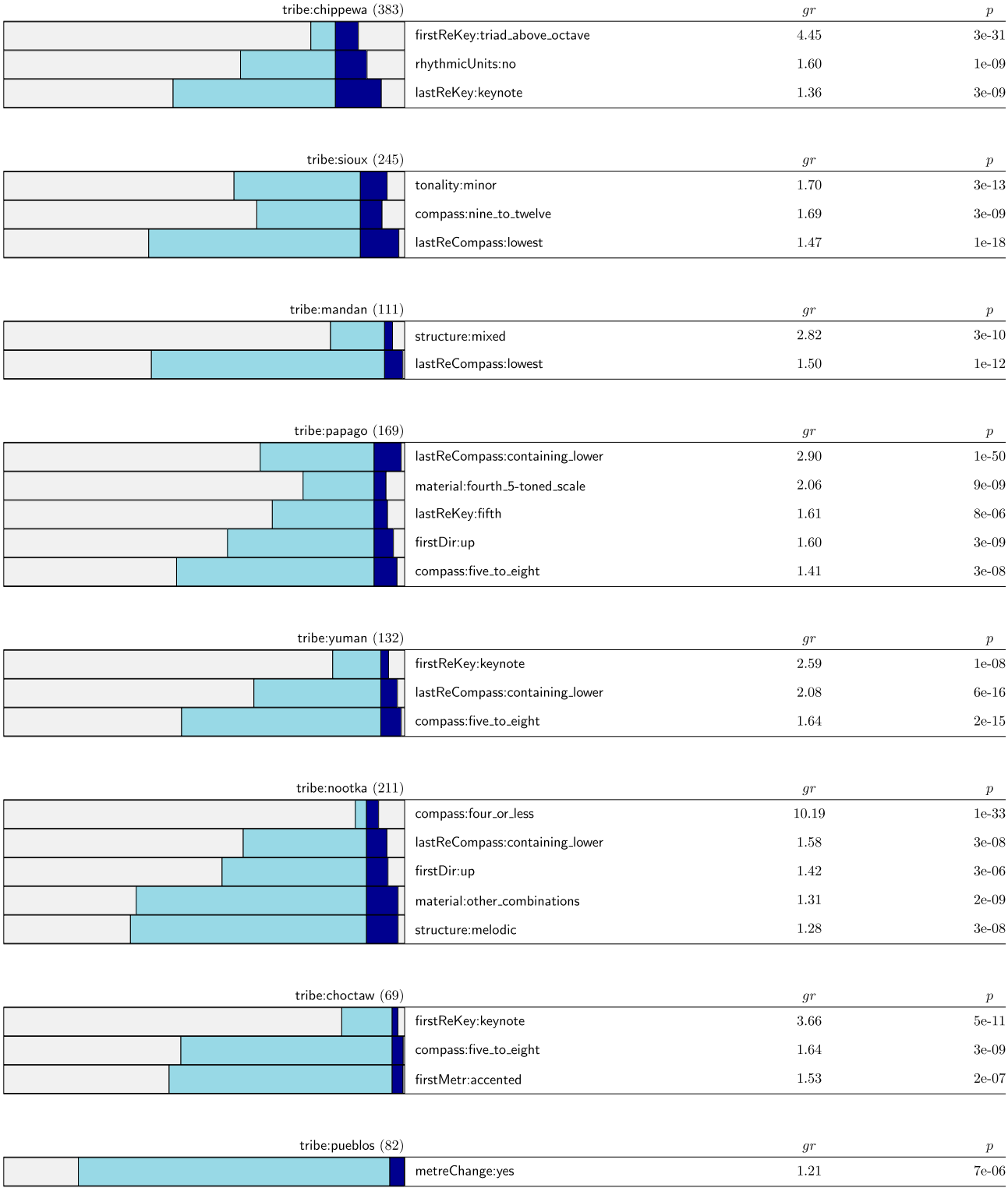


Figure 4. Interesting single-feature patterns for selected tribes, corresponding to features highlighted by Densmore (minimum growth rate $\theta = 1.2$, significance level $\alpha = 0.01$).

1943a, p. 183). The additional patterns based on jSymbolic attributes in Table 7—DirectionofMotion: High and MelArcDur: Low—also are over-represented in tribes with a high proportion of dance songs, in addition to the Choctaw: Pueblos songs (86% dance songs), and Yuman and Yaqui songs (55% dance songs). In fact, if the pattern discovery method is applied to the Densmore collection grouped by song types (Shanahan, Neubarth,

Table 7. Comparison of computational pattern discovery against Densmore's analysis. The table indicates growth rate and significance level of discovered patterns for mining of the cumulative data-set vs the complete data-set. Top: patterns which on the complete corpus are distinctive but not significant ($\theta = 1.4, \alpha = 0.01$). Middle: patterns which on the complete corpus are not distinctive ($\theta = 1.1$). Bottom: additional patterns.

Group	Pattern	Cumulative analysis	Complete corpus
tribe : northern_ute	lastReKey : fifth	$1.68^{\dagger}$	1.51
tribe : mandan	lastReKey : fifth	$1.65^{\dagger}$	$1.66^{\dagger}$
tribe : pawnee	structure : harmonic	1.41	1.61
tribe : pawnee	metreChange : no	1.75	1.42
tribe : yuman	firstDir : up	$1.61^{\ddagger}$	$1.46^{\dagger}$
tribe : yuman	metreChange : no	1.60	1.42
tribe : nootka	lastReKey : third	1.73	1.69
tribe : pueblos	rhythmicUnits : no	$1.72^{\dagger}$	$1.72^{\dagger}$
tribe : papago	structure : melodic	1.23	1.09
tribe : papago	initMetre : duple	1.16	1.06
tribe : nootka	lastReKey : second	$50.63^{\ddagger}$	$56.23^{\ddagger}$
tribe : yuman	DirectionofMotion : High	$2.44^{\ddagger}$	$1.87^{\ddagger}$
tribe : yuman	MelArcDur : Low	$1.63^{\ddagger}$	$1.46^{\ddagger}$
tribe : choctaw	DirectionofMotion : High	$1.70^{\ddagger}$	$1.62^{\ddagger}$
tribe : pueblos	MelArcDur : Low	$1.63^{\ddagger}$	$1.59^{\ddagger}$
tribe : pueblos	DirectionofMotion : High	$1.56^{\ddagger}$	$1.49^{\ddagger}$

Note: †: significant at $\alpha = 0.1$; ‡: significant at $\alpha = 0.01$.

& Conklin, 2016), DirectionofMotion : High is more significantly related to dances than to Choctaw or Pueblos songs. In the Yuman and Yaqui sample, on the other hand, both DirectionofMotion : High and MelArcDur : Low are also found in songs outside dances (90% and 83%, respectively, of Yuman and Yaqui songs). While Native American songs more generally have a largely descending melodic contour, Yuman songs show a more 'balanced upward and downward movement', creating 'level or undulating and sometimes arc-shaped' contours (Nettl, 1954, p. 302) of shorter duration (MelArcDur : Low). For Pueblos songs, oscillating contours (MelArcDur : Low) have also been described without specific reference to dance songs: particularly in the Western Pueblos style—represented by Acoma, Cochiti and Zuñi songs which constitute the majority (78%) of Pueblos songs in the corpus—'the direction of movement changes very frequently' (Herzog, 1936, p. 291).

4.2. Discovered feature-set patterns: Musical areas

Musical traits shared by several tribes are captured in Nettl's concept of musical areas (see Section 2.2.2). Example patterns in the Densmore corpus are presented in Figure 5, which displays interesting maximally general patterns for musical area groups (minimum growth rate $\theta = 2.0$, significance level $\alpha = 0.01$). Several of the discovered patterns can be directly related to Nettl's characterisation of the areas; others open questions for further investigation.

In the Eskimo—Northwest Coast area, many songs exhibit a recitative-like character, often within a 'restricted' range (Nettl, 1954, p. 54)—even less than a fifth, {compass : four_or_less}—and with 'usually undulating' melodic motion (Nettl, 1954, p. 54), captured by an over-representation of relatively short and narrow melodic arcs in combination with a high proportion of stepwise motion: {MelArcDur : Low , MelArcSize : Low , StepwiseMotion : High}. '[A]ctual pitches are relatively uncertain', although pitches are more established in songs of the Northwest-Coast tribes such as Nootka and Quileute (Nettl, 1954, p. 50). Still, Densmore found in her analyses that 'Nootka and Quileute songs do not adapt themselves entirely to the former bases of classification' regarding their tone material (Densmore, 1939, p. 42), resulting in a relatively high proportion of songs described as being irregular in tonality, {tonality : irregular}, or containing tone combinations other than pentatonic or heptatonic scales, with songs beginning on a tone within the octave above the keynote, due to their restricted compass, {material : other_combinations , firstReKey : other_within_octave}.

In contrast, songs of the Plains—Pueblo area tend to have wide to very wide ranges, with average ranges of an octave or tenth (Nettl, 1954). The melodic movement covering this range 'is also distinctive. It is primarily descending. Melodies show the following phrase pattern: each phrase descends, and each begins somewhat lower than the previous one', forming a terrace-type contour (Nettl, 1954, p. 351). Correspondingly, Plains—Pueblos songs in the Densmore collection more often than songs

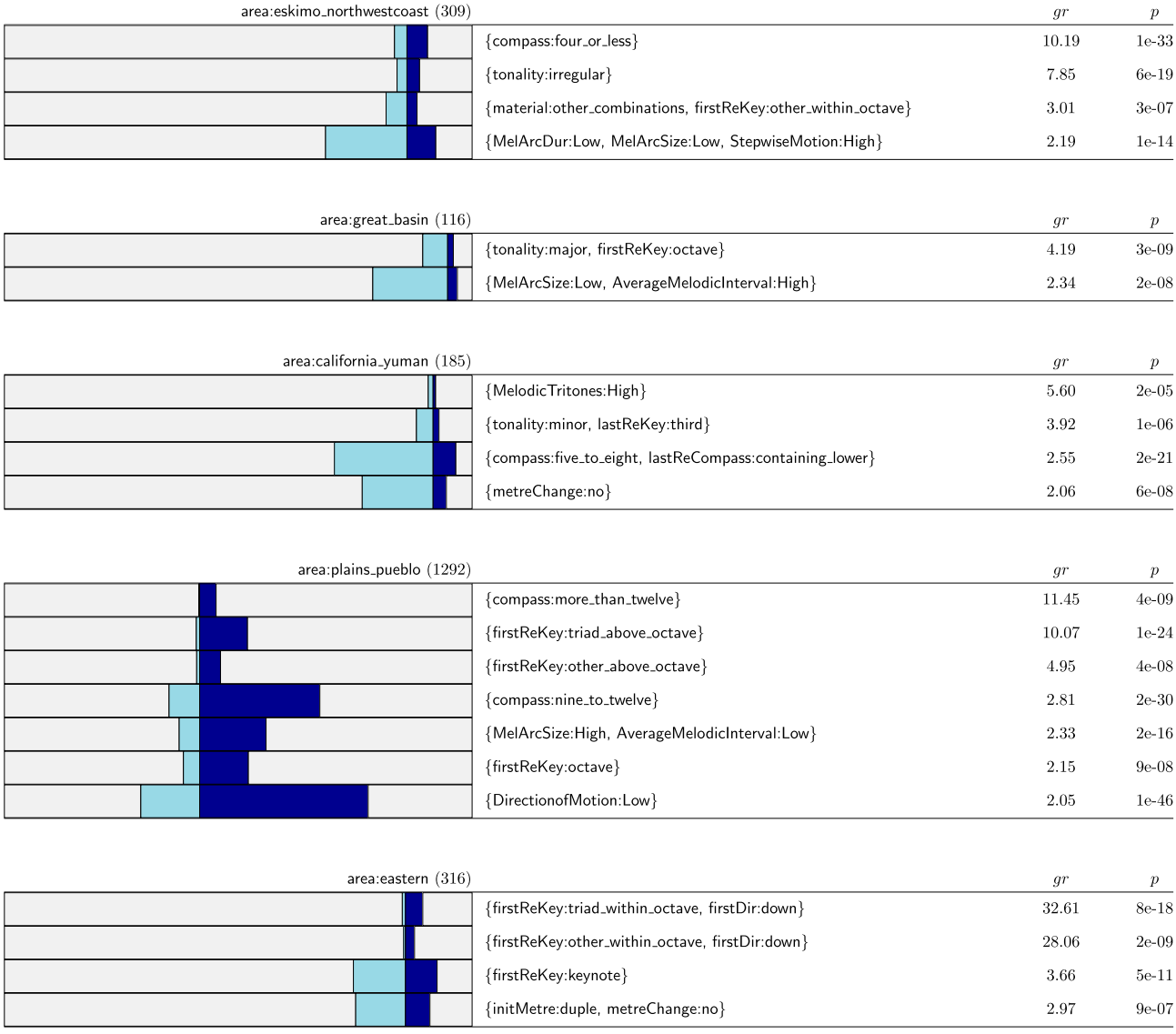


Figure 5. Discovered maximally general patterns for musical areas, ranked by growth rate (minimum growth rate $\theta = 2.0$, significance level $\alpha = 0.01$).

in the other areas begin an octave or higher above the keynote (patterns {firstReKey : triad_above_octave}, {firstReKey : other_above_octave} and {firstReKey : octave}) and descend (pattern {DirectionofMotion : Low}) through a wide range (patterns {compass : more_than_twelve} and {compass : nine_to_twelve}) in melodic arcs of relatively large ambitus filled by relatively small melodic intervals (pattern {MelArcSize : High , AverageMelodicInterval : Low}).

In comparison, the 'ranges of the Eastern area are relatively smaller than those of the Plains—Pueblo area [...]. In the Southeast the average is about a sixth' (Nettl, 1954, p. 360); Eastern songs in the corpus characteristically begin on the keynote, {firstReKey : keynote}, or on the third, fifth or another tone less than an octave above the keynote, followed by a downward motion, {firstReKey : triad_within_octave , firstDir : down} or {firstReKey : other_within_octave , firstDir : down}. The metrical pattern {initMetre : duple , metreChange : no} is particularly supported by Choctaw songs. Similar to the beginning on an accented tone (Section 4.1), Densmore relates these metrical characteristics to the high number of dances among the Choctaw songs: 'Double time is preferred by the Choctaw for the beginning of their songs, 83 percent having the first measure in 2–4 time. This would be expected, as a majority of recorded Choctaw songs are dance melodies' (Densmore, 1943a, p. 183). In terms of metre changes, Choctaw songs exhibit 'the smallest percentage in the songs under analysis and we note again that a majority of the recorded Choctaw songs are connected

with dances' (Densmore, 1943a, p. 183). The computational analysis reveals that the combination of these two features is almost threefold over-represented in the Eastern area compared to other areas. In the Densmore corpus duple metre, without metre changes, is indeed distinctive for dances (growth rate 1.46), but its over-representation in dances is not statistically significant.

Similarly, in the California—Yuman style songs without a change in metre, {metreChange : no}, are found more frequently than in most of the other areas: 'Isometric organization is somewhat more common in this area than elsewhere on the continent (with isolated exceptions, such as the Salish and the Southeastern U.S.), but is not found in the majority of the songs' (Nettl, 1954, p. 303). In the Densmore corpus, the feature is supported by 34% of the songs in the area, compared to 17% in the background (growth rate 2.59). The range of Yuman and Californian songs tends to be between a fifth and ninth or tenth (Herzog, 1928; Nettl, 1954), melodies often moving downwards towards the keynote; '[in] some songs the downward flow does not stop when it reaches the tonic, but is carried below it' (Herzog, 1936, p. 302). California—Yuman songs in the Densmore collection seem to support such observations: 59% of California—Yuman songs satisfy the pattern {compass : five_to_eight , lastReCompass : containing_lower}, against only 23% of songs in other areas. Comparing Pima musical style against the style of Pueblo tribes, Herzog (1936) occasionally draws parallels to characteristics of the Yuman, Plateau Shoshonean and some Californian tribes: in both Pima and Pueblo songs he observes 'the occurrence of the augmented fourth' (p. 309). In the Densmore corpus, melodic tritones are discovered as distinctive for the California—Yuman area (pattern {MelodicTritone : High}, growth rate 5.60).

The reported patterns for the Great Basin area are less easily related to observations by Nettl. Fewer interesting ($\theta = 2.0$, $\alpha = 0.01$) patterns are discovered for this area. Potentially this reflects the impression that in many respects, e.g. a 'lack of specialized melodic movement' (Nettl, 1954, pp. 297–298), the Great Basin style seems less delineated than other style areas. In the Densmore corpus this is compounded by the fact that the area is represented by only one tribe, the Northern Ute, which Nettl characterises as 'to some extent marginal to this area' (Nettl, 1954, p. 297), sharing traits of the Great Basin and of the Plains styles. The beginning on the octave might point towards the Plains—Pueblo area; however, while for Great Basin songs this beginning occurs particularly in songs of major tonality, {firstReKey : octave , tonality : major}, in Plains—Pueblo songs it is the combination with minor tonality which is statistically distinctive (growth rate 6.93, p-value 4e-13). On the other hand, the perhaps most characteristic feature of the Great Basin style—the paired-phrase pattern, each phrase being repeated (Herzog, 1935a), often in combination with 'the tendency for most phrases to end on the tonic' (Nettl, 1954, p. 298)—would not be captured by the features analysed in this study. More generally, the current Humdrum encoding of the Densmore corpus does not contain phrase annotations, nor do Densmore's transcriptions. If phrase information was available, characteristics not modelled in the current analysis, such as phrase patterns, could also be studied.

5. Discussion and conclusions

Computational and quantitative studies of folk music go back to the early days of exploring computational methods in musicology (e.g. Bronson, 1959; Suchoff, 1970). More recently, digitisation of music corpora and advances in data mining have led to a renewed interest in collection-level and comparative music analysis, not only of folk music (e.g. Cook, 2004; Temperley & Van Handel, 2013). Supervised descriptive mining methods find interesting patterns in group-labelled data: in the current study, contrast set mining was adapted to discover global-feature patterns which are over-represented in a tribe or musical area of Native American music compared to other tribes or areas, evaluated by distinctiveness and statistical significance. Results can be presented as associations between groups and sets of attribute—value pairs which capture distinguishing characteristics of tribes and musical areas. In this respect supervised pattern discovery differs from attribute selection which identifies attributes that contribute to separating groups, but does not reveal characteristic attribute values.

Different from predictive modelling, for descriptive pattern discovery there are no established global evaluation measures, such as error rate in classification or recall in information retrieval; rather measures assess the potential interestingness of individual patterns, such as their distinctiveness for a given group. Densmore's collection of Native American music offers a valuable opportunity to compare results of computational pattern discovery against Densmore's own analyses and additional ethnomusicological studies. Moreover, Densmore's analysis, while not explicitly calculating pattern interestingness, in its approach essentially represents a manual application of supervised descriptive pattern discovery. This paper has applied computational pattern discovery to the Densmore corpus, which was organised into 15 tribe groups. Representing songs by global features extracted from Densmore's tabular analyses, the computational analysis discovers patterns for tribes which correspond to characteristics also observed by Densmore. When both

Densmore features and computationally extracted features are used to characterise musical style areas in Native American music, many discovered patterns can be related to the original ethnomusicological description of these areas (Nettl, 1954). Further supported by a simulation which yields very few false discoveries on the data-set randomised through computational permutation, these findings suggest that supervised descriptive discovery methods can indeed identify global-feature patterns which are interesting both statistically and musicologically.

Applying contrast set mining extends Densmore's comparison of tribes using single features to analysing sets of global features. In this, the method presented in this paper not only goes beyond the earlier manual analyses of e.g. Densmore, but also more recent computational studies which test single features (e.g. Neubarth, Johnson, & Conklin, 2013) or fixed-size sets of features, in particular pairs of features (e.g. Taminau et al., 2009). In contrast, our method supports the discovery of flexible-size feature sets by mining for maximally general interesting patterns. The concept of maximally general distinctive patterns has previously been developed for sequential patterns of varying length (Conklin, 2010). Here we extend the notion of maximally general patterns in two ways: to global-feature patterns and to further interestingness measures. More specifically, following contrast set mining methods, we consider distinctiveness and statistical significance in the evaluation of candidate patterns. An additional contribution beyond earlier descriptive mining studies, motivated specifically by the Densmore corpus, is the explicit consideration of missing values: without this the evaluation of associations is inaccurate.

While computational methods provide flexibility over manual analysis e.g. with respect to discovering multi-dimensional patterns (feature-set in addition to single-feature patterns), considering additional features extracted from the digitised songs (here jSymbolic features in addition to Densmore features), generalising discovered patterns for different data-sets (cumulative analysis and complete corpus) and partitioning the corpus in different ways (here tribe groups and area groups), in other respects some flexibility is lost: in observations drawn from her tabular analyses, at times Densmore aggregates features in varying ways, for example: 'In the Ute songs the initial tones of the songs, in about 75 per cent [sic], are either the keynote, its third, fifth or octave [...]. In the Chippewa and Sioux songs the preference is for the twelfth and fifth' (Densmore, 1922, p. 52). Representation and discovery of such aggregated patterns would require the use of disjunction in pattern descriptions (Loekito & Bailey, 2006), which would, however, not only increase computational complexity but also lead to less concise patterns.

In this paper, we have shown how supervised descriptive data mining can identify interesting and comprehensible patterns in a large music collection, which support and complement musicological observations. Computational methods can be used to test hypotheses based on existing analyses or to explore data-sets and suggest patterns for further investigation. Descriptive data mining methods thus have an important role in corpus-level music analysis.

Acknowledgements

The authors would like to thank Olivia Barrow, Eva Shanahan, Craig Sapp, Paul Von Hippel, and David Huron for encoding work on the data-set at various stages.

Funding

This research was partially supported by the project Lrn2Cre8 which is funded by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET [grant number 610859].

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases*. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC (pp. 207–216).

Atzmüller, M. (2015). Subgroup discovery - advanced review. *WIREs Data Mining and Knowledge Discovery, 5*, 35–49.

Bay, S. D., & Pazzani, M. J. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery, 5*(3), 213–246.

Bronson, B. (1959). Towards the comparative analysis of British-American folk tunes. *The Journal of American Folklore, 72*(284), 165–191.

Browner, T. (2000). Making and singing pow-wow songs: Text, form and the significance of culture-based analysis. *Ethnomusicology, 44*(2), 214–233.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis, 14*, 547–554.

Conklin, D. (2013a). Antipattern discovery in folk tunes. *Journal of New Music Research, 42*(2), 161–169.

Conklin, D. (2013b). Multiple viewpoint systems for music classification. *Journal of New Music Research, 42*(1), 19–26.

Conklin, D., & Anagnostopoulou, C. (2011). Comparative pattern analysis of Cretan folk songs. *Journal of New Music Research, 40*(2), 119–125.

Cook, N. (2004). Computational and comparative musicology. In E. Clarke & N. Cook (Eds.), *Empirical musicology: Aims, methods, prospects* (pp. 103–126). Oxford: Oxford University Press.

Densmore, F. (1910). Chippewa music. *Bureau of American Ethnology, Bulletin 45*. Washington, DC: Smithsonian Institution.
Densmore, F. (1913). Chippewa music II. *Bureau of American Ethnology, Bulletin 53*. Washington, DC: Smithsonian Institution.
Densmore, F. (1915). The study of Indian music. *The Musical Quarterly, 1*(2), 187–197.
Densmore, F. (1918). Teton Sioux music. *Bureau of American Ethnology, Bulletin 61*. Washington, DC: Smithsonian Institution.
Densmore, F. (1922). Northern Ute music. *Bureau of American Ethnology, Bulletin 75*. Washington, DC: Smithsonian Institution.
Densmore, F. (1923). Mandan and Hidatsa music. *Bureau of American Ethnology, Bulletin 80*. Washington, DC: Smithsonian Institution.
Densmore, F. (1929a). Papago music. *Bureau of American Ethnology, Bulletin 90*. Washington, DC: Smithsonian Institution.
Densmore, F. (1929b). Pawnee music. *Bureau of American Ethnology, Bulletin 93*. Washington, DC: Smithsonian Institution.
Densmore, F. (1932a). Menominee music. *Bureau of American Ethnology, Bulletin 102*. Washington, DC: Smithsonian Institution.
Densmore, F. (1932b). Yuman and Yaqui music. *Bureau of American Ethnology, Bulletin 110*. Washington, DC: Smithsonian Institution.
Densmore, F. (1936). *Cheyenne and Arapaho music*. Los Angeles, CA: Southwest Museum.
Densmore, F. (1939). Nootka and Quileute music. *Bureau of American Ethnology, Bulletin 124*. Washington, DC: Smithsonian Institution.
Densmore, F. (1943a). Choctaw music. *Bureau of American Ethnology, Bulletin 136*. Washington, DC: Smithsonian Institution.
Densmore, F. (1943b). Music of the Indians of British Columbia. *Bureau of American Ethnology, Bulletin 136*. Washington, DC: Smithsonian Institution.
Densmore, F. (1956). Seminole music. *Bureau of American Ethnology, Bulletin 161*. Washington, DC: Smithsonian Institution.
Densmore, F. (1957). Music of Acoma, Isleta, Cochiti and Zuñi Pueblos. *Bureau of American Ethnology, Bulletin 165*. Washington, DC: Smithsonian Institution.
Densmore, F. (1958). *Music of the Maidu Indians of California*. Los Angeles, CA: Southwest Museum.
Dong, G., & Li, J. (1999). *Efficient mining of emerging patterns: discovering trends and differences*. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA (pp. 43–52).
Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: a survey. *ACM Computing Surveys, 38*(3), 1–32.
Gundlach, R. H. (1932). A quantitative analysis of Indian music. *The American Journal of Psychology, 44*(1), 133–145.
Hatton, O. T. (1986). In the tradition: Grass dance musical style and female pow-wow singers. *Ethnomusicology, 30*(2), 197–222.
Herzog, G. (1928). The Yuman musical style. *The Journal of American Folklore, 41*(160), 183–231.
Herzog, G. (1935a). Plains ghost dance and great basin music. *American Anthropologist, 37*, 403–419.
Herzog, G. (1935b). Special song types in North American Indian music. *Zeitschrift für vergleichende Musikwissenschaft, 3*, 23–33.
Herzog, G. (1936). A comparison of Pueblo and Pima musical styles. *The Journal of American Folklore, 49*(194), 283–417.
Hillewaere, R. (2013). Computational models for folk music classification (Ph.D. thesis). Brussels: Vrije Universiteit Brussel.
Hofmann, C. (1946). Frances Densmore and the music of the American Indian. *The Journal of American Folklore, 59*(231), 45–50.
Klösgen, W. (1996). Explora: a multipattern and multistrategy discovery assistant. In U. Fayyad, G. Piatetsky-Shapiro, & P. Smyth (Eds.), *Advances in knowledge discovery and data mining* (pp. 249–271). Cambridge, MA and London: MIT Press.
Kurath, G. (1953). Native choreographic areas of North America. *American Anthropologist, 55*(1), 60–73.
Levine, V. L. (1998). American Indian musics, past and present. In D. Nicholls (Ed.), *The Cambridge history of American music* (pp. 3–29). Cambridge: Cambridge University Press.
Liu, G., Zhang, H., & Wong, L. (2011). Controlling false positives in association rule mining. *Proceedings of the VLDB Endowment, 5*(2), 145–156.
Loekito, E., & Bailey, J. (2006). *Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams*. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, PA (pp. 307–316).
Martins, M. L., & Silla, C. N. (2015). *Irish traditional ethnomusicology analysis using decision trees and high level symbolic features*. 12th Sound and Music Computing Conference (SMC), Maynooth. w/o pages.
McKay, C. (2010). Automatic music classification with jMIR (Ph.D. thesis). Montreal: McGill University.
Nettl, B. (1954). North American Indian musical styles. *The Journal of American Folklore, 67*(263,265,266), 44–56, 297–307, 351–368.
Nettl, B. (1955). Musical culture of the Arapaho. *The Musical Quarterly, 41*(3), 325–331.
Nettl, B. (1967). Studies in Blackfoot Indian musical culture, part I: traditional uses and functions. *Ethnomusicology, 11*(2), 141–160.
Nettl, B., & Blum, S. (1968). Studies in Blackfoot Indian musical culture, part III: three genres of song. *Ethnomusicology, 12*(1), 11–48.
Neubarth, K., Johnson, C. G., & Conklin, D. (2013). *Discovery of mediating association rules for folk music analysis*. 5th International Workshop on Music and Machine Learning at ECML/PKDD 2013 (MML, 2013), Prague. w/o pages.
Nguyen, D., Luo, W., Phung, D., & Venkatesh, S. (2016). *Exceptional contrast set mining: moving beyond the deluge of the obvious*. Proceedings of the 29th Australasian Joint Conference on Advances in Artificial Intelligence, Hobart (pp. 455–468).
Novak, P. K., Lavrač, N., & Webb, G. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set,

emerging pattern and subgroup mining. *Journal of Machine Learning Research, 10*, 377–403.
Novak, P. K., Lavrač, N., & Webb, G. I. (2010). Supervised descriptive rule induction. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 938–941). New York, NY: Springer.
Shanahan, D., Neubarth, K., & Conklin, D. (2016). *Mining musical traits of social functions in Native American music*. Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016), New York, NY (pp. 681–687).
Shanahan, D., & Shanahan, E. (2014). *The Densmore collection of Native American songs: A new corpus for studies of effects of geography and social function in music*. Proceedings of the 13th International Conference for Music Perception and Cognition (ICMPC 2014), Seoul (pp. 206–209).
Smithsonian Institution (1971). List of publications of the Bureau of American Ethnology. *Bureau of American Ethnology, Bulletin 200*. Washington, DC: Smithsonian Institution.
Suchoff, B. (1970). Computer-oriented comparative musicology. In H. Lincoln (Ed.), *The computer and music* (pp. 193–206). Ithaca, NY and London: Cornell University Press.
Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowé, A., & Manderick, B. (2009). *Descriptive subgroup mining of folk music*. 2nd International Workshop on Machine Learning and Music at ECML/PKDD 2009 (MML 2009), Bled. w/o pages.
Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson.
Temperley, D., & Van Handel, L. (2013). Introduction to the special issues on corpus methods. *Music Perception, 31*(1), 1–3.
van Kranenburg, P., Volk, A., & Wiering, F. (2013). A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research, 42*(1), 1–18.
Vennum, T., Jr (2000). Locating the Seri on the musical map of Indian North America. *Journal of the Southwest, 42*(3), 635–760.
Volk, A., & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae, 16*(3), 317–339.
Webb, G. I. (2007). Discovering significant patterns. *Machine Learning, 68*(1), 1–33.
Webb, G. I., Butler, S., & Newlands, D. (2003). *On detecting differences between groups*. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), Washington, DC (pp. 256–265).
Wu, J., He, Z., Gu, F., Liu, X., Zhou, J., & Yang, C. (2016). Computing exact permutation p-values for association rules. *Information Sciences, 346–347*, 146–162.

Appendix 1. Fisher's exact test for associations

The one-tailed Fisher's exact test can be used to determine the p-value of an association $X \rightarrow G$. Different from a χ^2 test, Fisher's test can also be reliably applied if counts in the inner cells of the contingency table (Table 4) are small. For notational convenience, let $x = n(X, G)$, $n = n(G)$, and $m = n(X)$ (adjusted to account for missing values, as was described in Section 3.3). The probability of drawing m pieces from a total of N pieces, and finding exactly x pieces satisfying the group G is given by the hypergeometric distribution:

$$\mathbb{H}(x, n, m, N) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}.$$

The probability of finding at least x pieces satisfying the group G is then given by:

$$\sum_{i=x}^{m} \mathbb{H}(i, n, m, N).$$