DECADE Mini Symposium

November 10, 2017

FLASHPOINTS: MINING EXCEPTIONAL PAIRWISE BEHAVIOR IN VOTE OR RATING DATASETS

AUTHORS.

Adnene BELFODIL Philippe LAMARRE

Marc PLANTEVIT Sylvie CAZALENS

LIRIS INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON UCBL Lyon 1 cnrs

INTRODUCTION

Collaborative rating platforms

Movielens

Yelp

Voting datasets

European parliament voting

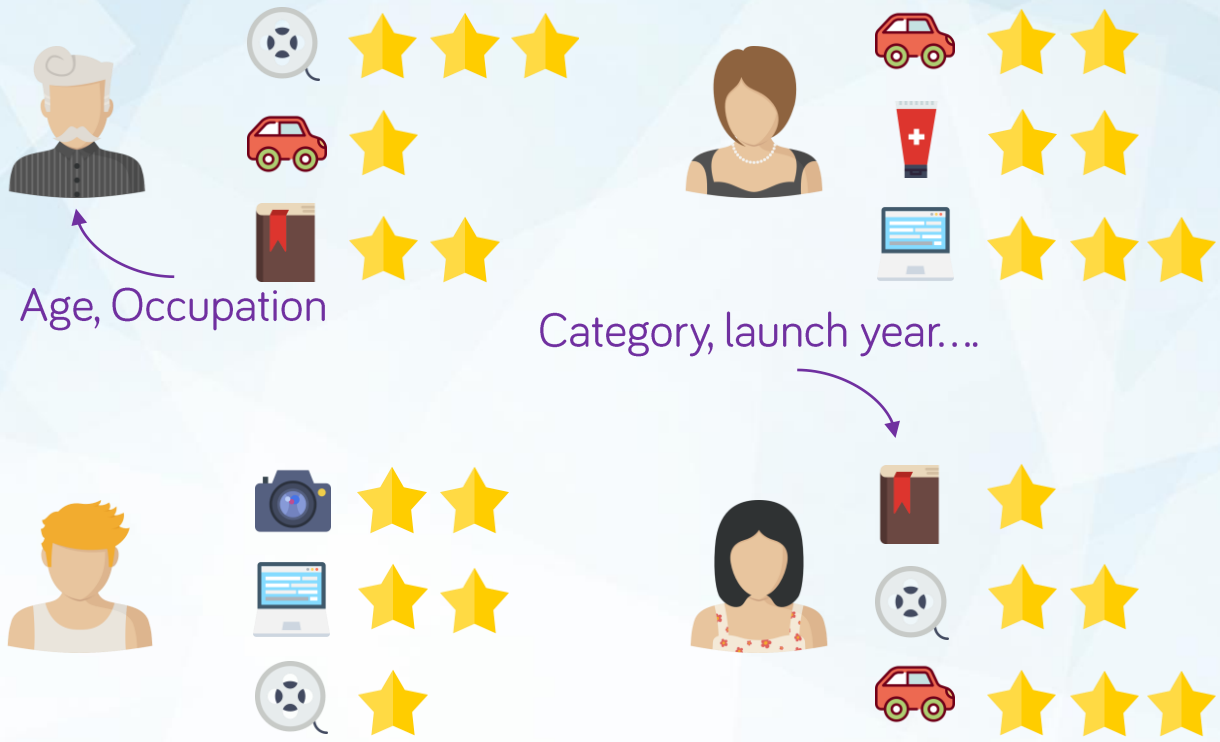
INTRODUCTION

Collaborative rating platforms

Movielens

Yelp

Collaborative rating platforms



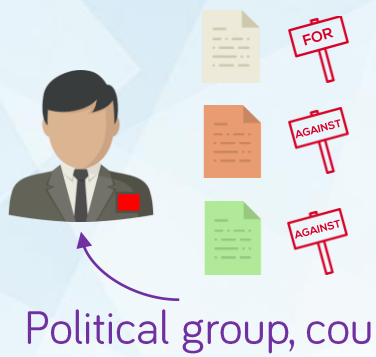
INTRODUCTION

DECADE
Belfodil, Lamarre, Cazalens & Plantevit

Voting datasets

European parliament voting

Voting datasets



Political group, country…

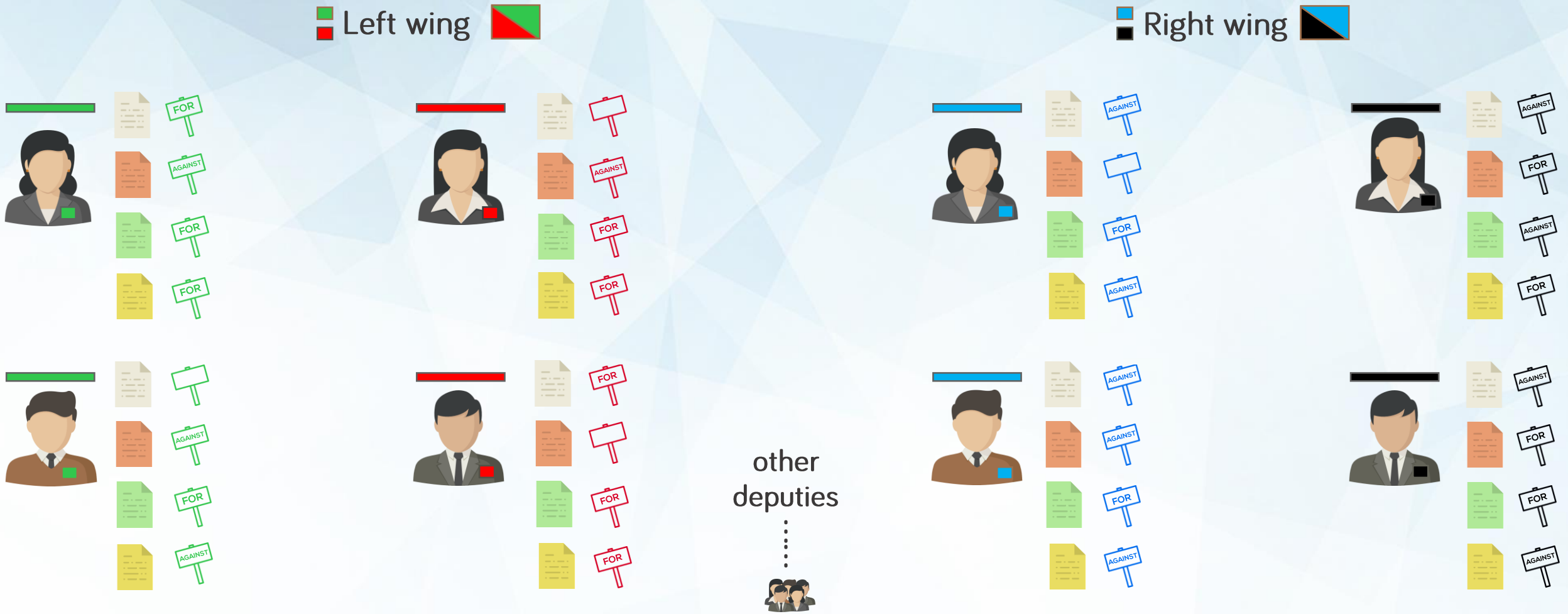
Theme, ballot date …

*context

other deputies

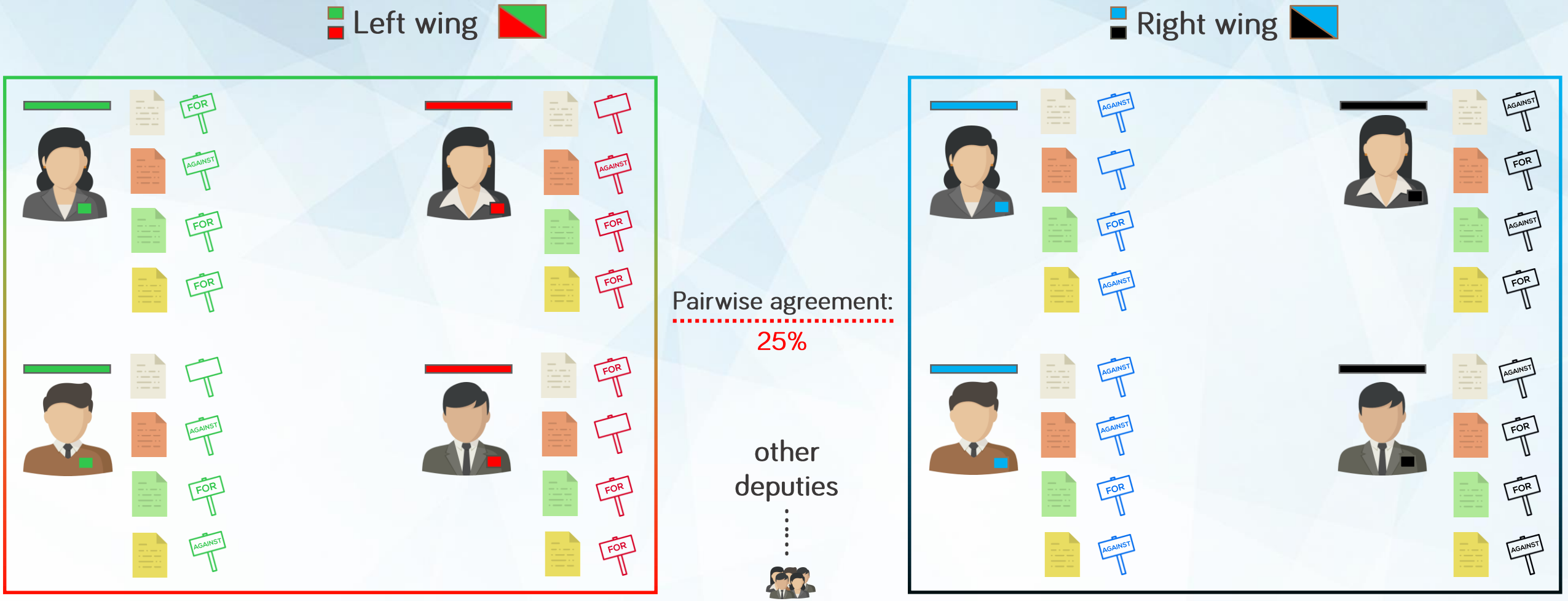
CONTENTCHECK

INTRODUCTION



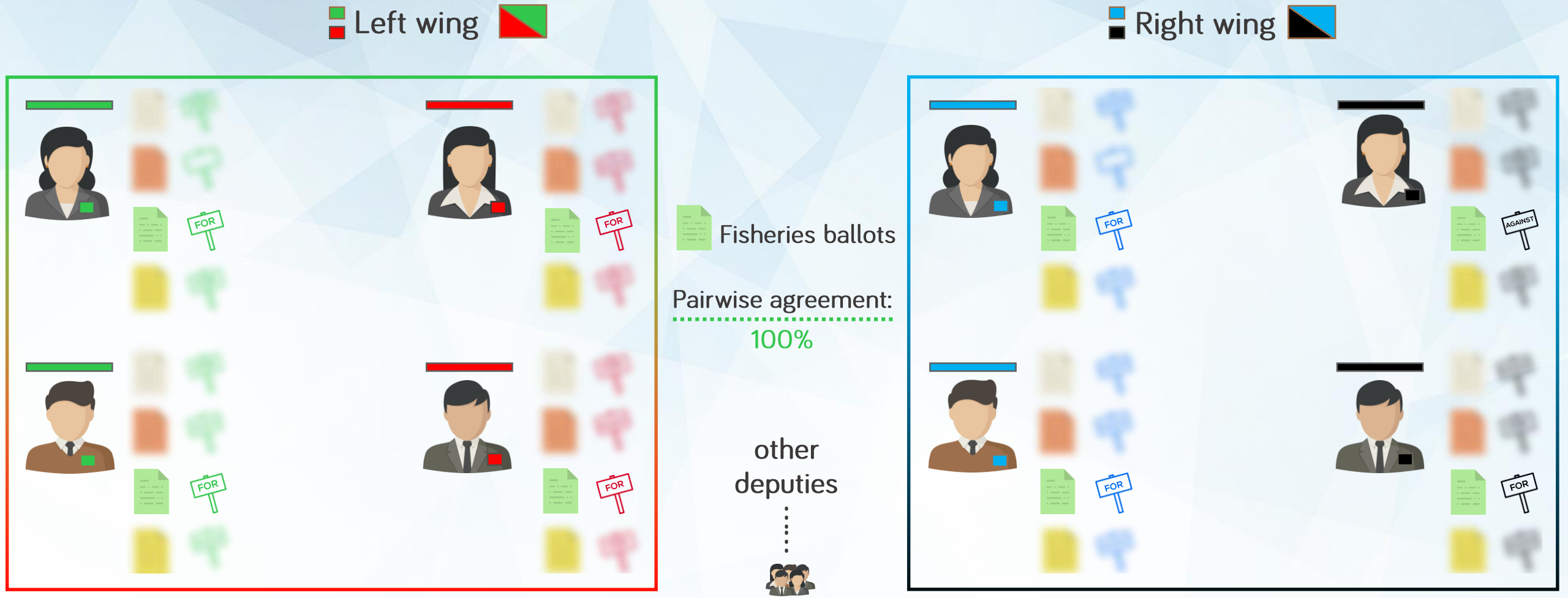
INTRODUCTION

DECADE
Belfodil, Lamarre, Cazalens & Plantevit



INTRODUCTION

DECADE
Belfodil, Lamarre, Cazalens & Plantevit



PROBLEM DEFINITION

We introduce the problem of discovering particular contexts and collections of individuals such that their pairwise behavior exceptionally differs from their usual one

PROBLEM DEFINITION

We introduce the problem of discovering particular contexts and collections of individuals such that their pairwise behavior exceptionally differs from their usual one

Find the top-k three-set patterns (c, g', g") w.r.t. some quality measure φ

PROBLEM DEFINITION

We introduce the problem of discovering particular contexts and collections of individuals such that their pairwise behavior exceptionally differs from their usual one

Find the top-k three-set patterns (c, g', g'') w.r.t. some quality measure φ

Context

definition by intent of a
subset of items

PROBLEM DEFINITION

We introduce the problem of discovering particular contexts and collections of individuals such that their pairwise behavior exceptionally differs from their usual one

Find the top-k three-set patterns (c, g', g") w.r.t. some quality measure φ

Context

definition by intent of a subset of items

g' & g"

definition by intent of a subset of individuals

PROBLEM DEFINITION

We introduce the problem of discovering particular contexts and collections of individuals such that their pairwise behavior exceptionally differs from their usual one

Find the top-k three-set patterns (c, g', g") w.r.t. some quality measure φ

Example:

(Consumer Protection in General Ballots voted in between 2015 and 2016 , German Deputies, Italian Deputies)

We observe a significant decrease of pairwise agreement

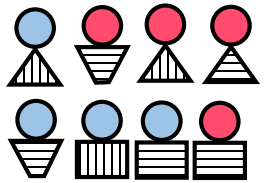
SUMMARY

- DSC (Discovering similarities change) Framework
- DSC algorithm
- Use cases

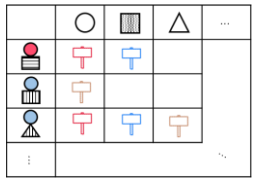
SUMMARY

- DSC (Discovering similarities change) Framework
- DSC algorithm
- Use cases

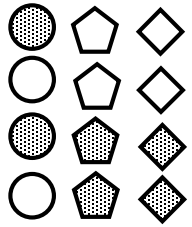
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



Reviewers
(eg. Users, Deputies)

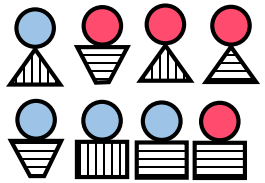


Reviews
(eg. Scores, Votes)

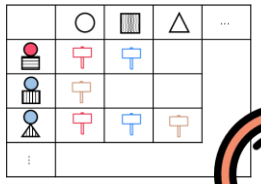


Reviewees
(eg. Movies, Vote ballots)

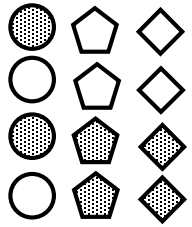
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



Reviewers
(eg. Users, Deputies)



Reviews
(eg. Scores, Votes)

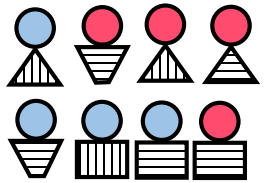


Reviewees
(eg. Movies, Vote ballots)

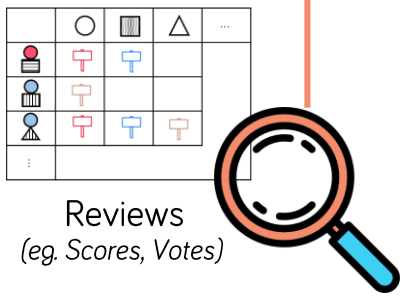
Dataset example: Parliament voting dataset

Items (Ballots) - E			*Individuals (Deputies) - U*			*Outcome*
Idsession	*Date*	*Theme*	*Full name*	*National Party*	*Political Group*	*Vote*
001	2017/03/17	1.10 Justice 2.10 Europe coop	Lavrilleux	LR	PPE	For
001	2017/03/17	1.10 Justice 2.10 Europe coop	Philippot	FN	ENF	Against
002	2017/04/11	3.10 Agriculture	Lavrilleux	LR	PPE	For
002	2017/04/11	3.10 Agriculture	Philippot	FN	ENF	For
002	2017/04/11	3.10 Agriculture	Arnatu	FN	ENF	For
003	2017/04/11	1.20 Security	Le Grip	LR	PPE	Abstain

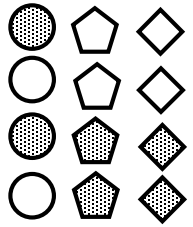
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



Reviewers
(eg. Users, Deputies)



Reviews
(eg. Scores, Votes)



Reviewees
(eg. Movies, Vote ballots)

Dataset example: Parliament voting dataset

*Items (Ballots) - **E***			*Individuals (Deputies) - **U***			*Outcome*
Idsession	*Date*	*Theme*	*Full name*	*National Party*	*Political Group*	*Vote*
001	2017/03/17	1.10 Justice 2.10 Europe coop	Lavrilleux	LR	PPE	For
001	2017/03/17	1.10 Justice 2.10 Europe coop	Philippot	FN	ENF	Against
002	2017/04/11	3.10 Agriculture	Lavrilleux	LR	PPE	For
002	2017/04/11	3.10 Agriculture	Philippot	FN	ENF	For
002	2017/04/11	3.10 Agriculture	Arnatu	FN	ENF	For
003	2017/04/11	1.20 Security	Le Grip	LR	PPE	Abstain

Descriptions attributes* over items (context)

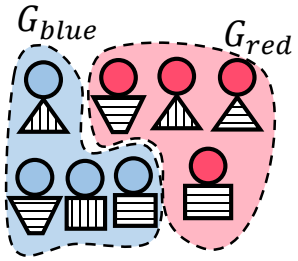
Descriptions attributes* over individuals

*numeric, nominal, hierarchical multi-tag attributes

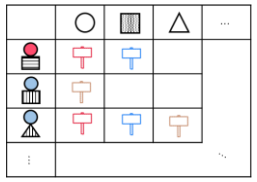
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK

DSC OVERVIEW

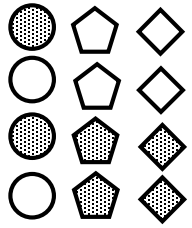
1 *Constitute groups*
(eg. By head color)



Reviewers
(eg. Users, Deputies)

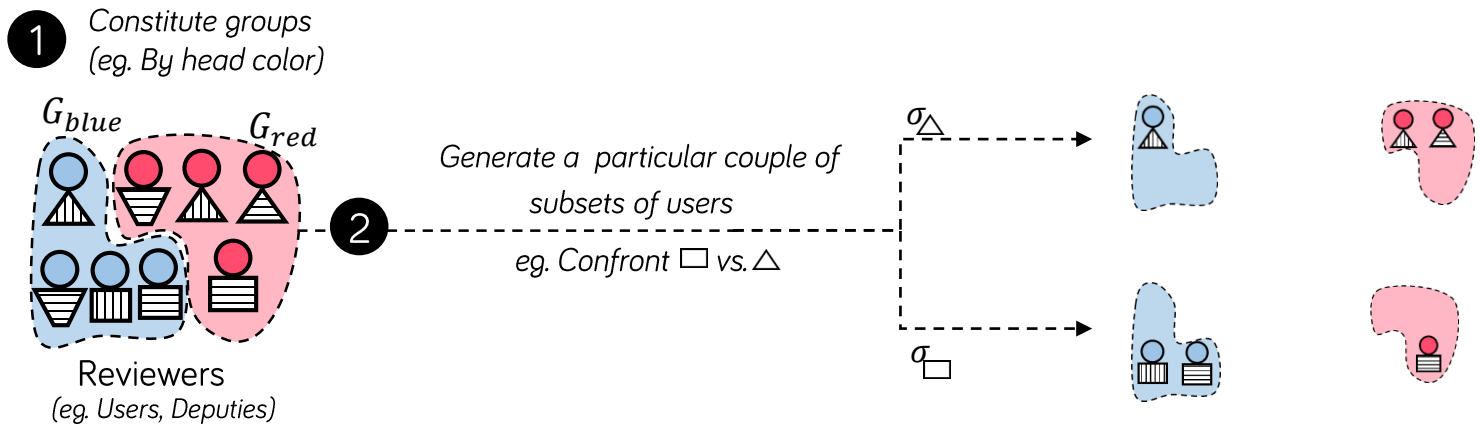


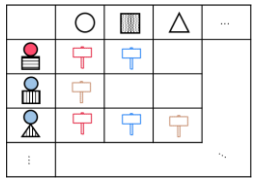
Reviews
(eg. Scores, Votes)



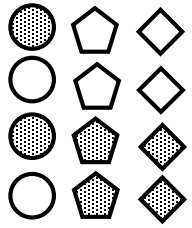
Reviewees
(eg. Movies, Vote ballots)

DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



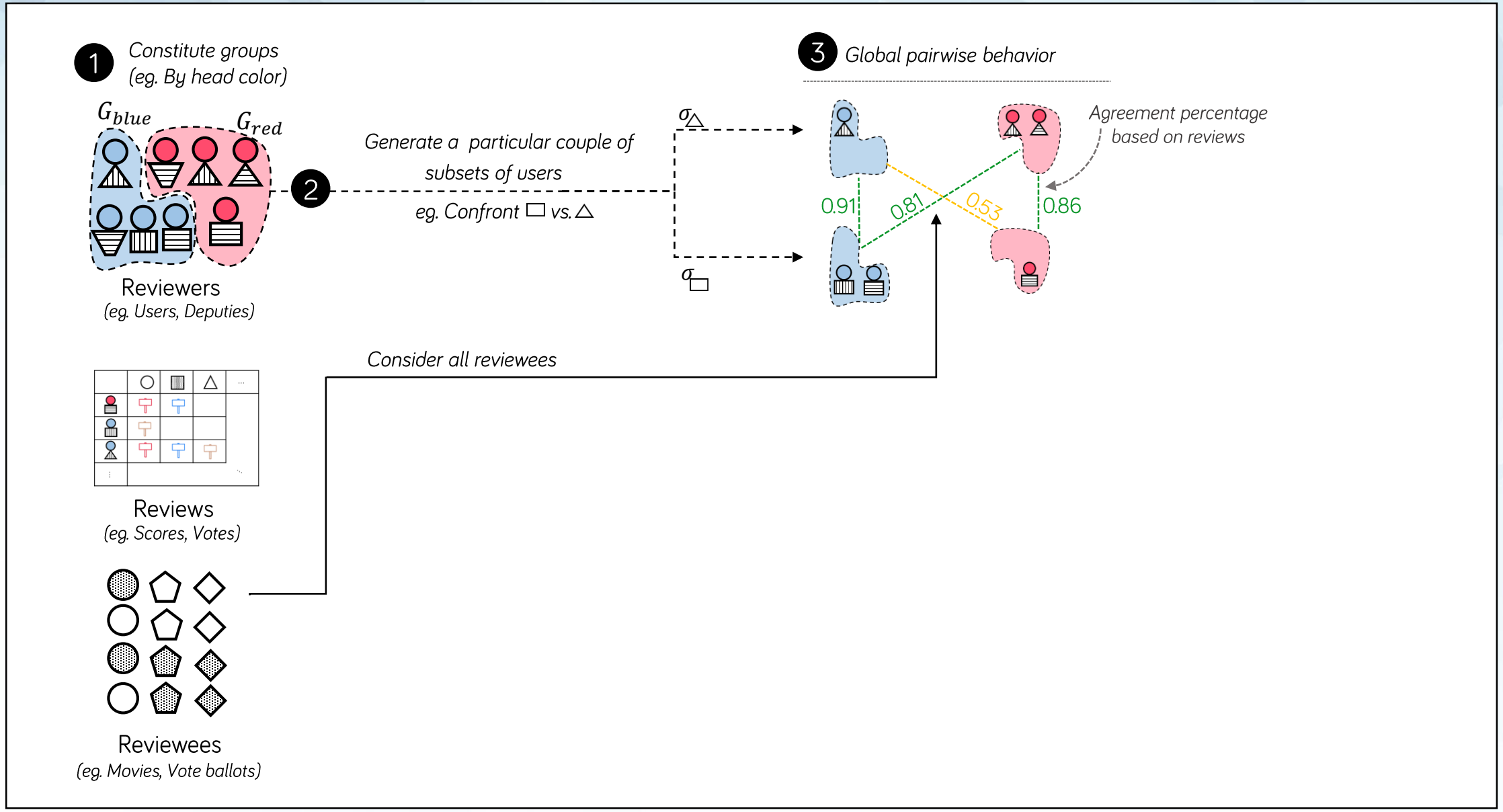


Reviews
(eg. Scores, Votes)

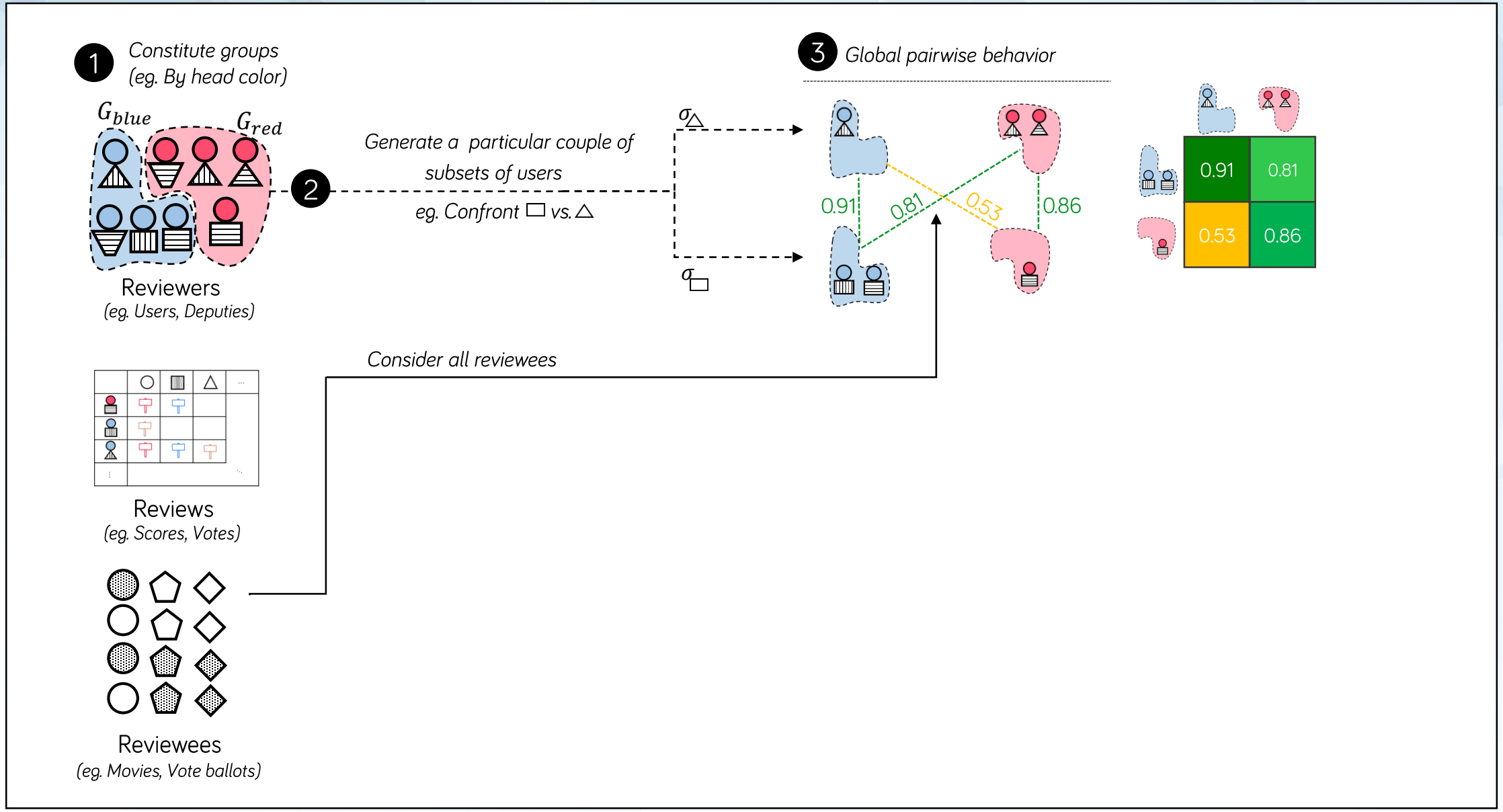


Reviewees
(eg. Movies, Vote ballots)

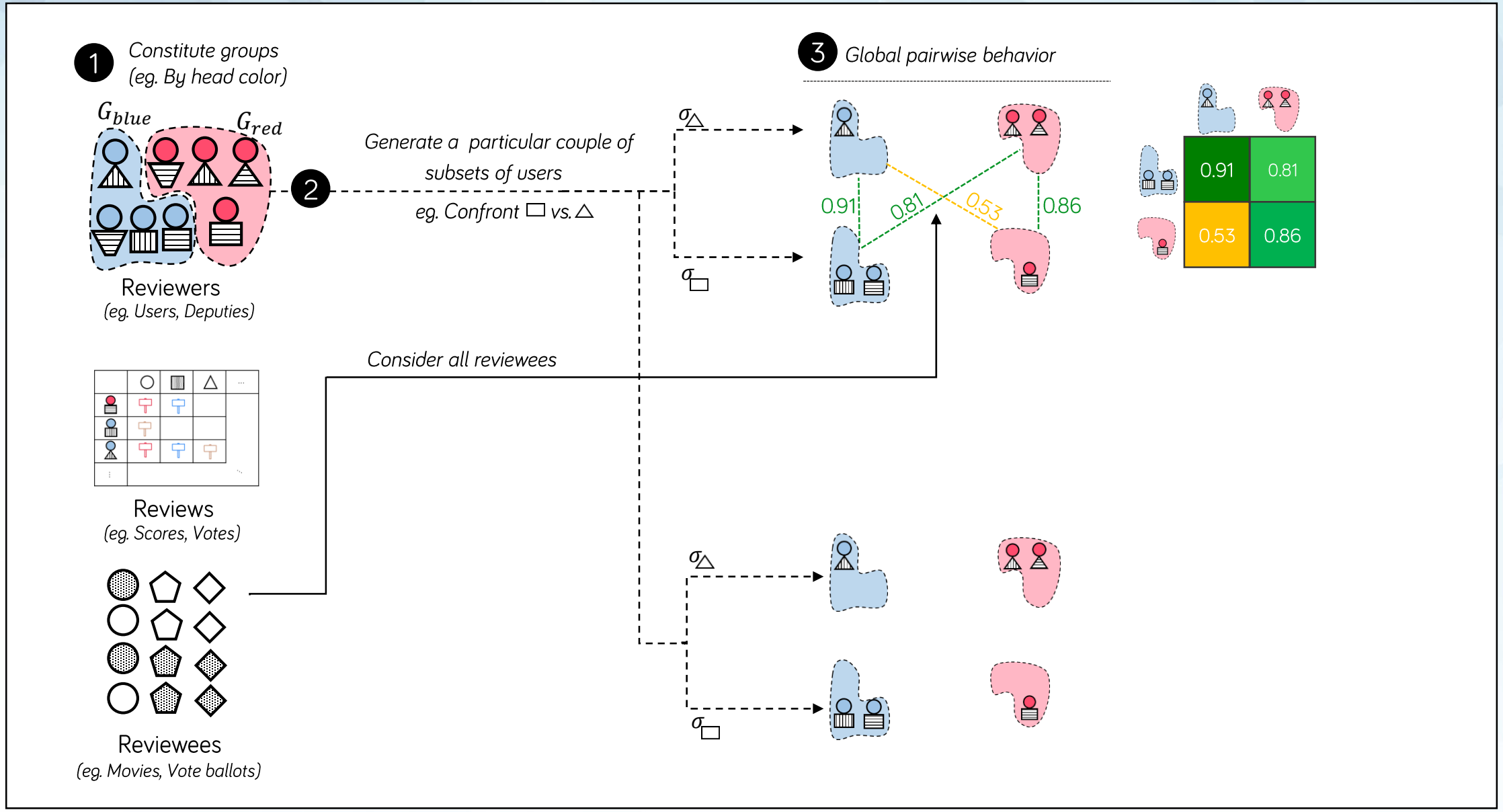
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



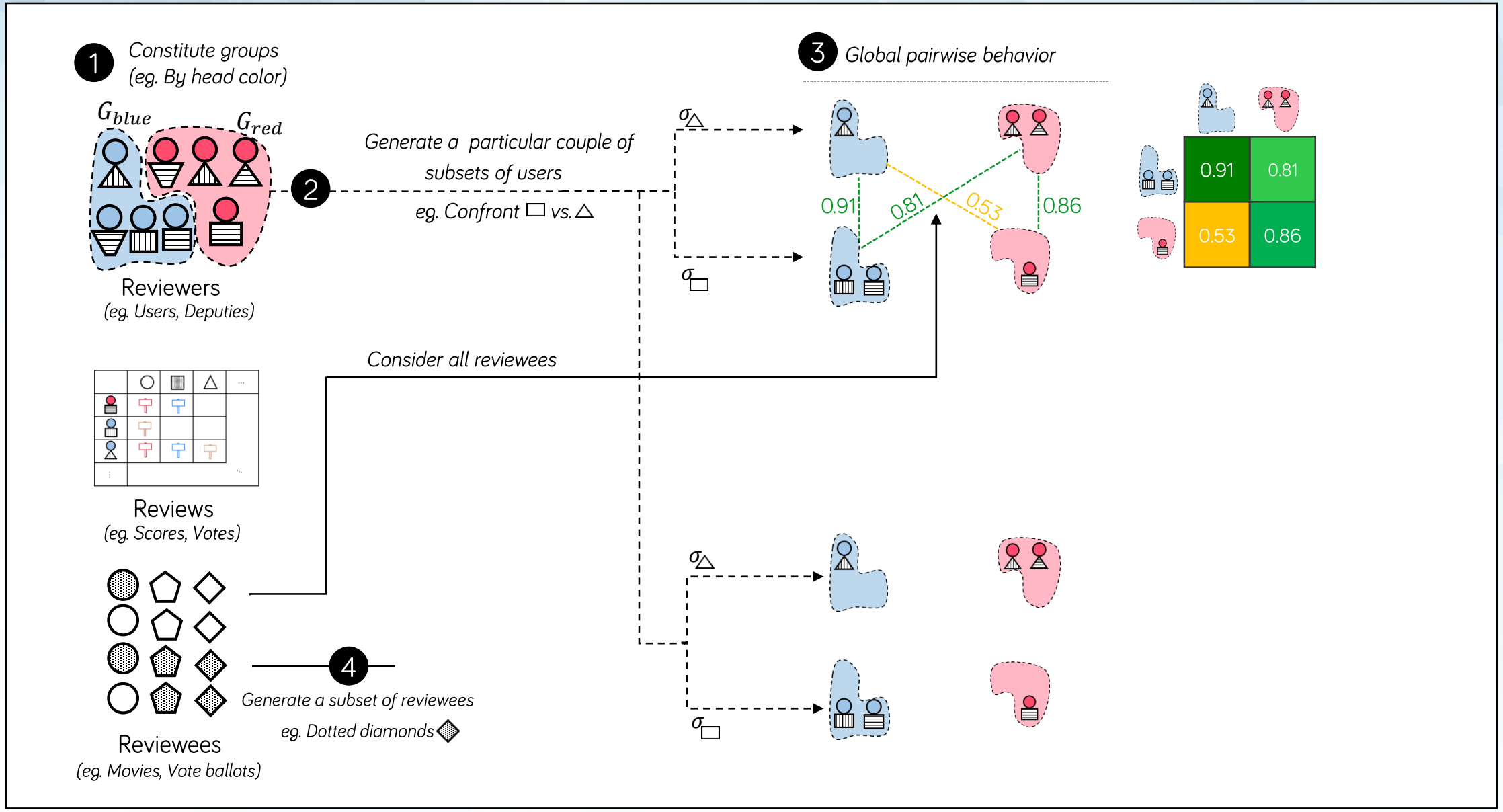
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



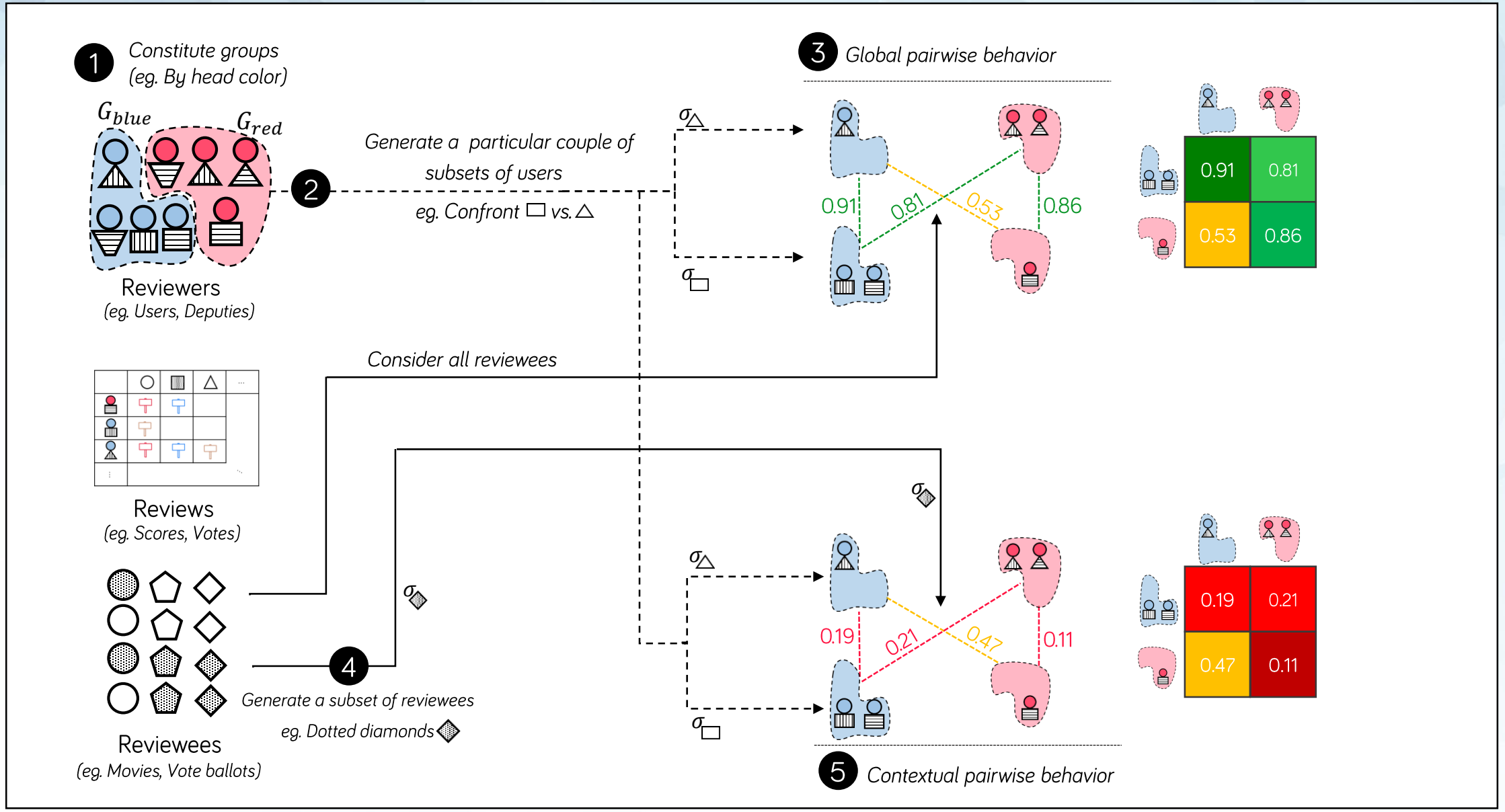
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



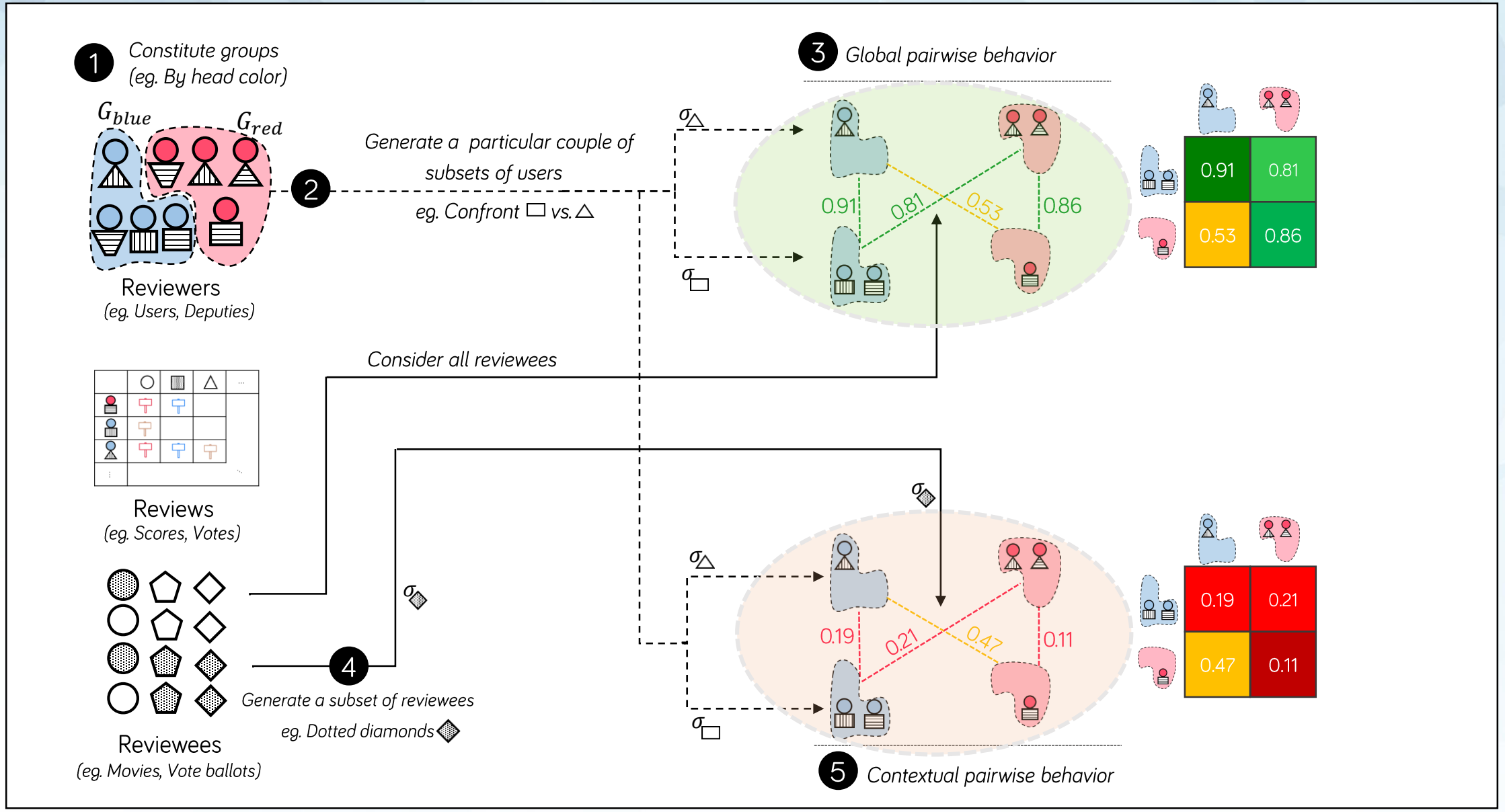
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



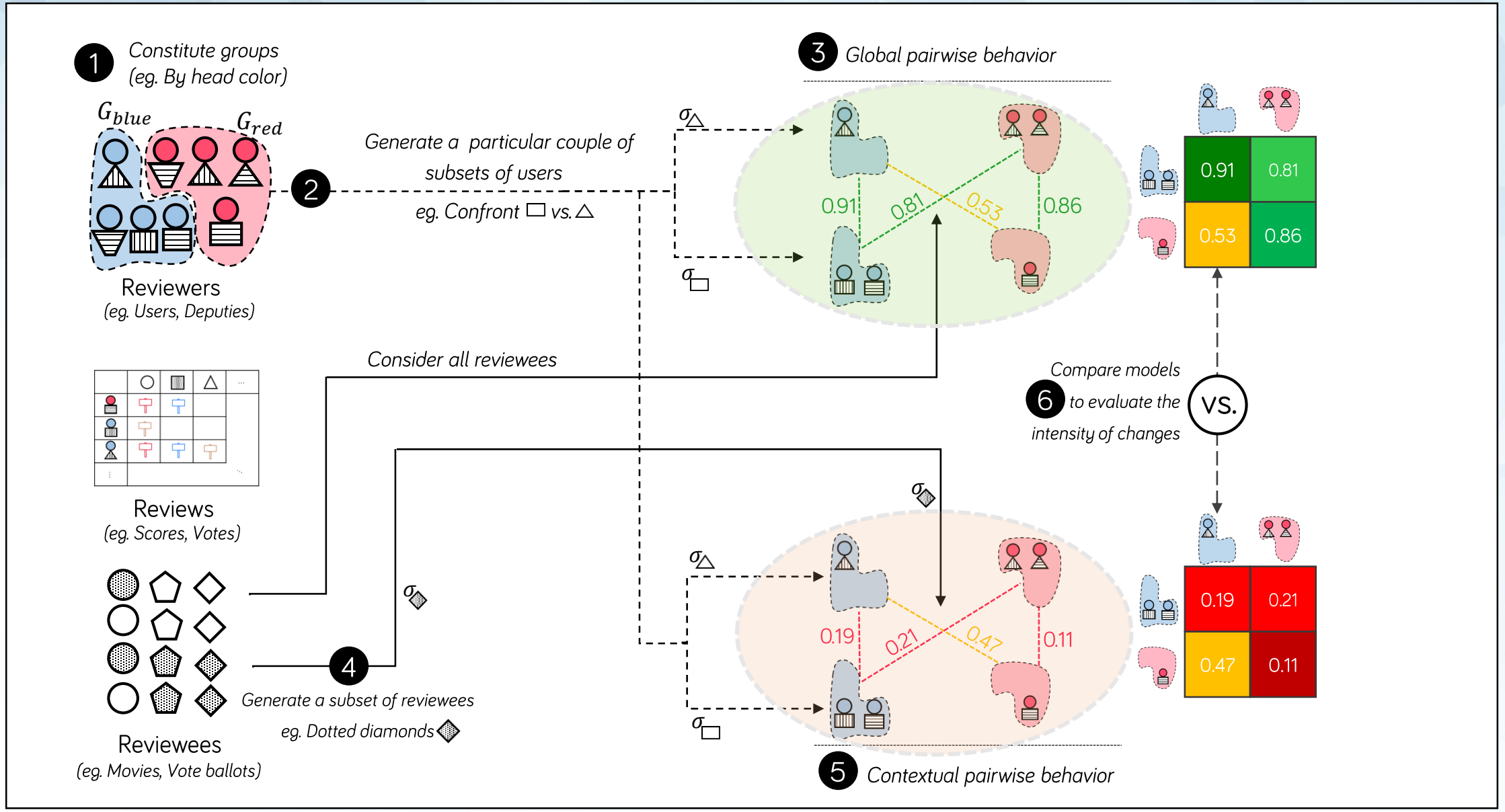
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



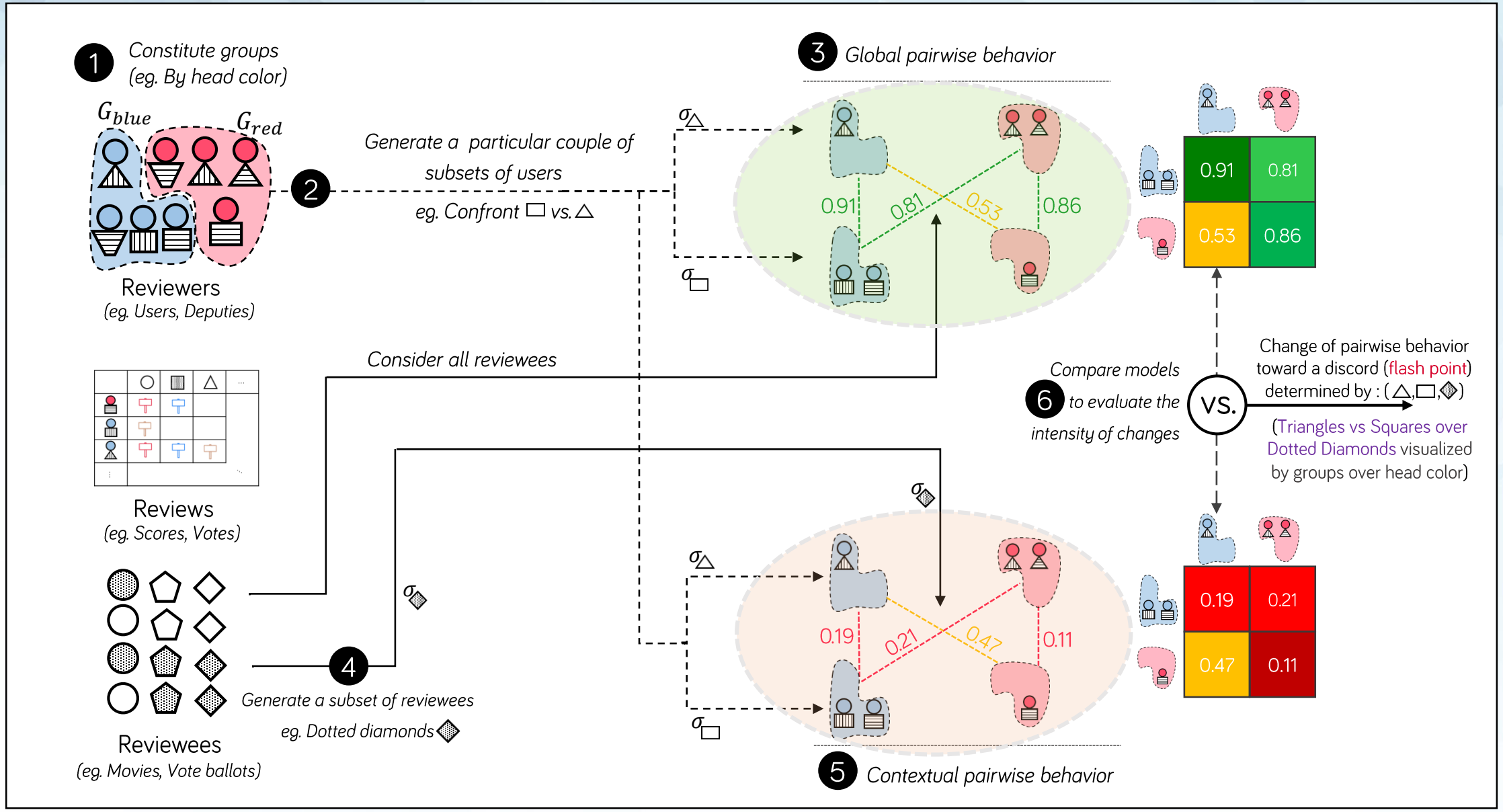
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



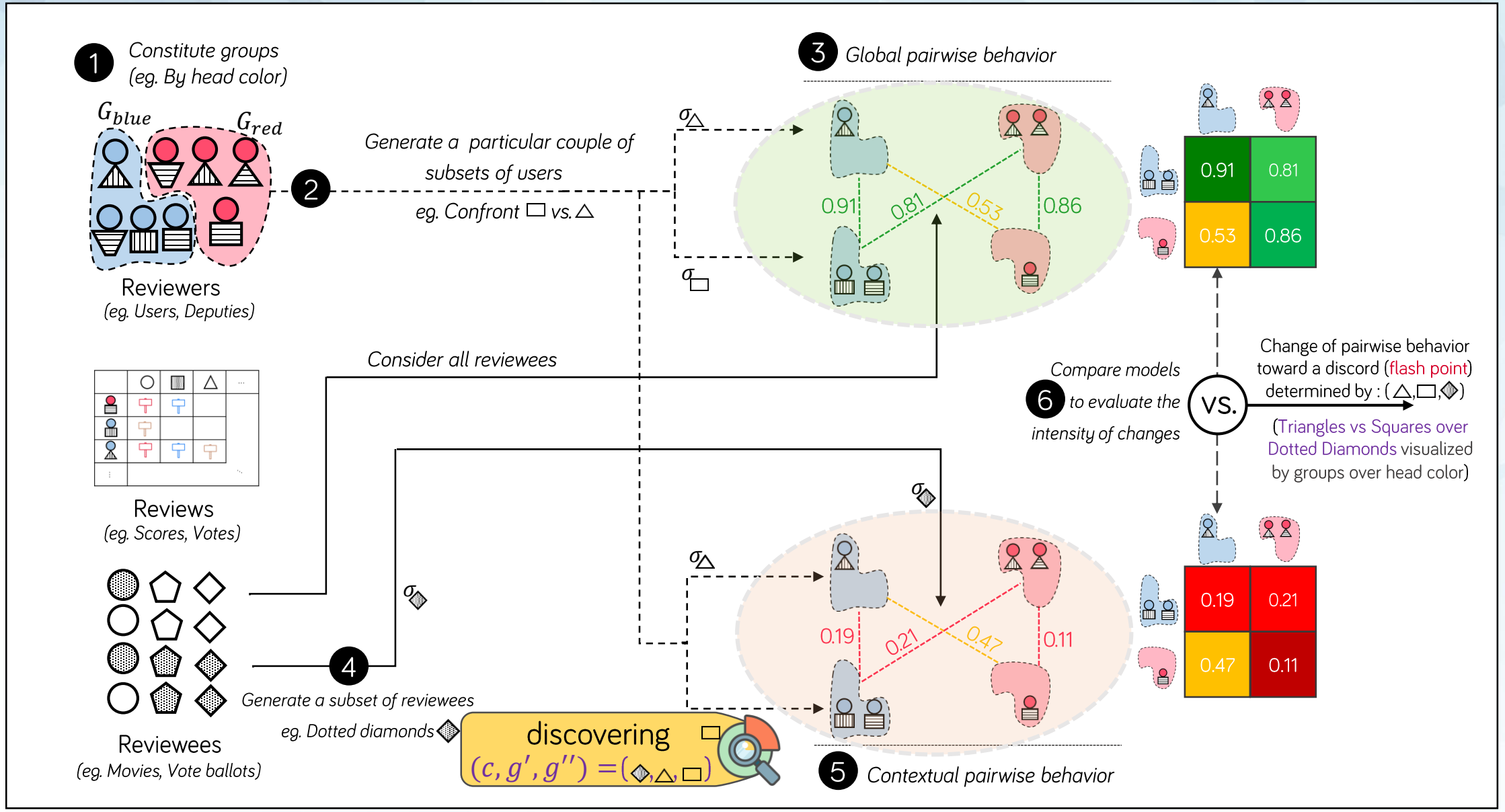
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK

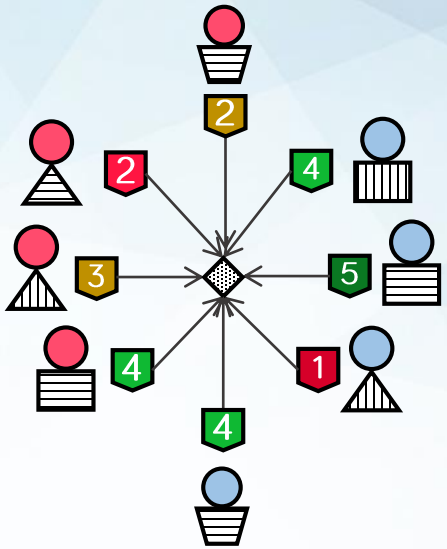


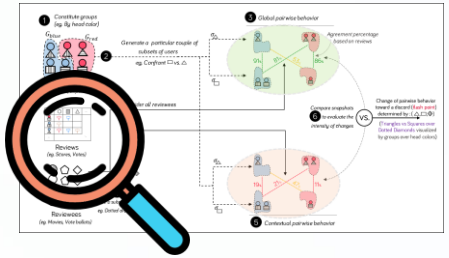
DISCOVERING SIMILARITIES CHANGE (DSC) FRAMEWORK



FROM INDIVIDUAL TO GROUPS BEHAVIOR

Individuals described by (color, shapes)

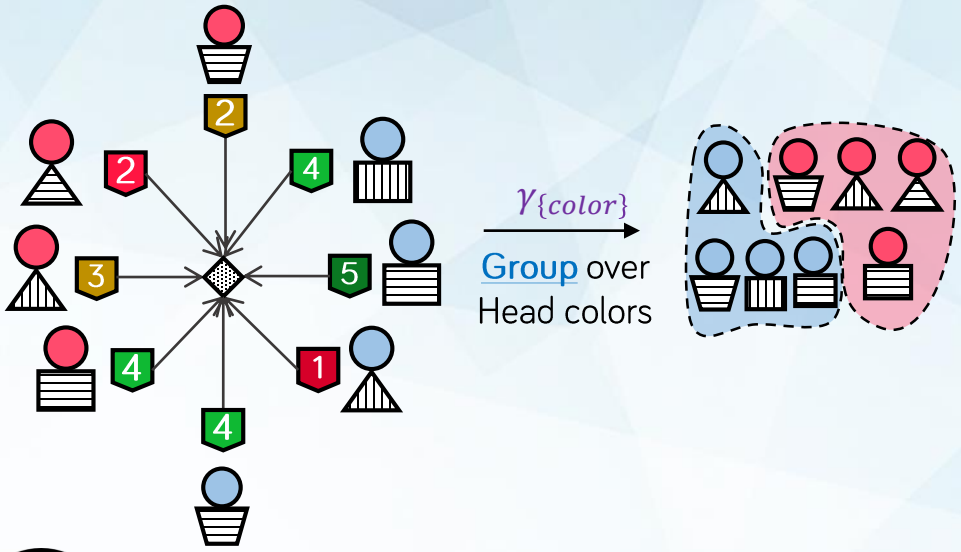


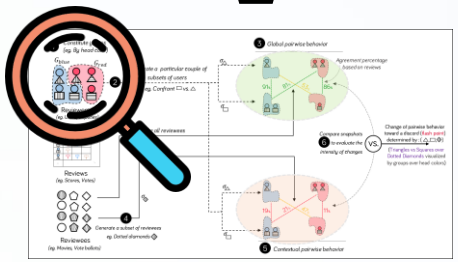


FROM INDIVIDUAL TO GROUPS BEHAVIOR

DECADE
Belfodil, Lamarre, Cazalens & Plantevit

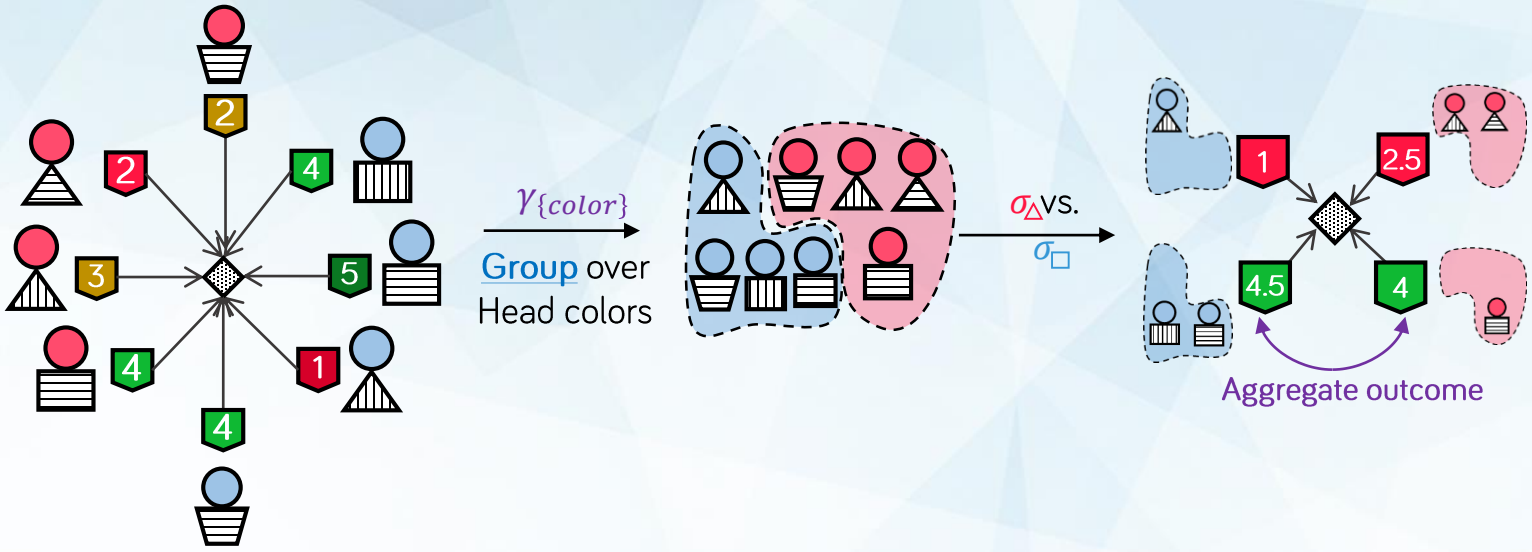
Individuals described by (color, shapes)

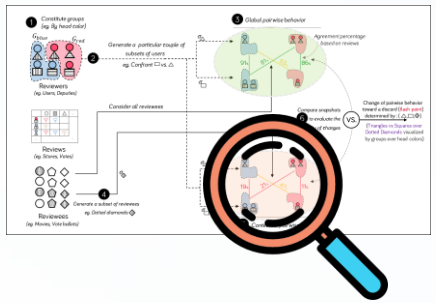




FROM INDIVIDUAL TO GROUPS BEHAVIOR

Individuals described by (color, shapes)





Several aggregation operators, e.g.:

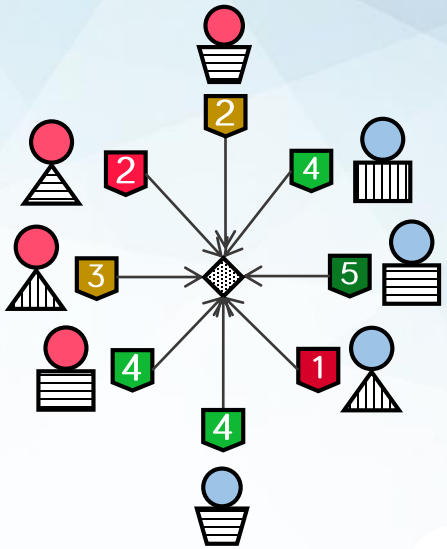
Majorities votes: $\theta(e,G) = argmax_{v \in O}\ count(v, \{outcome(e,u) | u \in G\})$

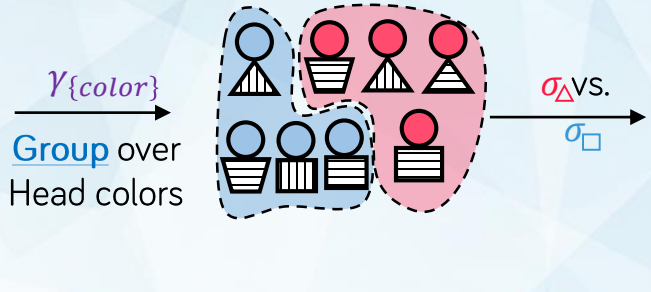
Rating average: $\theta(e,G) = \frac{1}{|G|} \sum_{u \in G} outcome(e,u)$

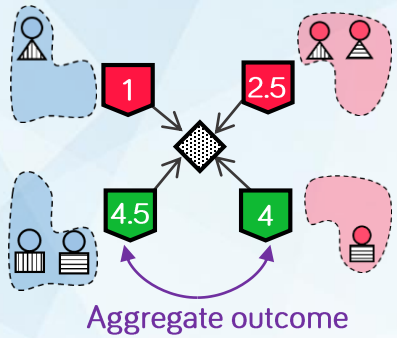
FROM INDIVIDUAL TO GROUPS BEHAVIOR

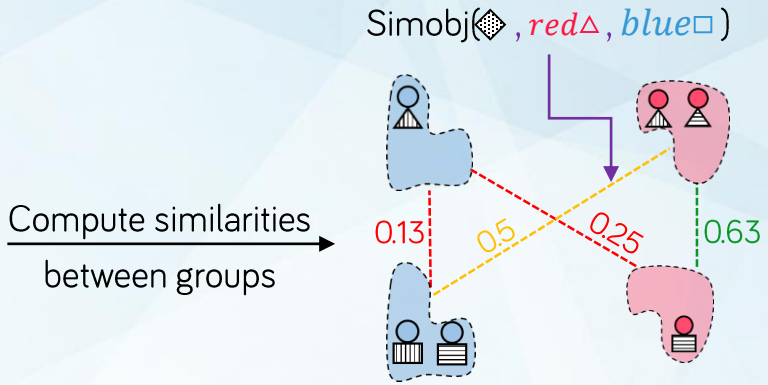
DECADE
Belfodil, Lamarre, Cazalens & Plantevit

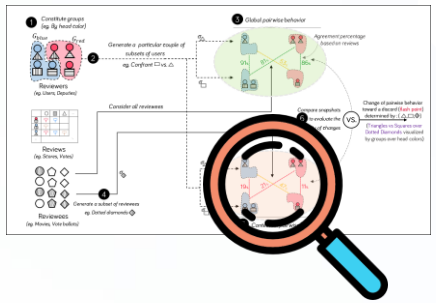
Individuals described by (color, shapes)











Similarity functions over one item: $simobj$

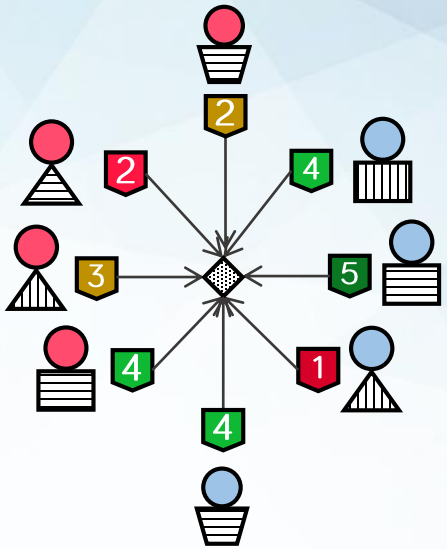
Rating agreement: $simobj(e, G_1, G_2) = 1 - \frac{1}{scaleSize} |\theta(e, G_1) - \theta(e, G_2)|$

Voting similarity: $simobj(e, G_1, G_2) = \delta_{\theta(e,G_1)\theta(e,G_2)}$

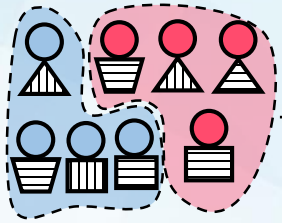
FROM INDIVIDUAL TO GROUPS BEHAVIOR

DECADE
Belfodil, Lamarre, Cazalens & Plantevit

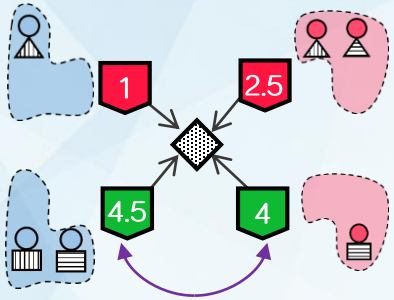
Individuals described by (color, shapes)



$\gamma_{\{color\}}$
Group over
Head colors

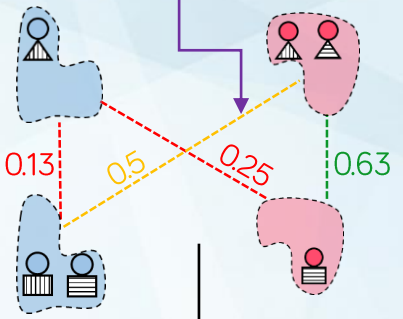


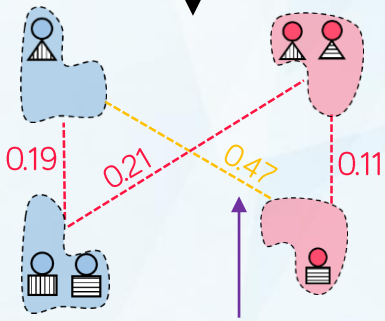
$\sigma_{\triangle}$ vs. $\sigma_{\square}$



Compute similarities
between groups

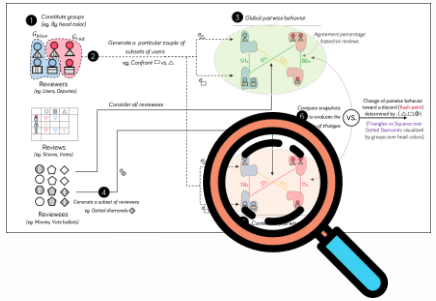
Simobj(, $red\triangle$, $blue\square$)





sim(, $red\triangle$, $blue\square$)

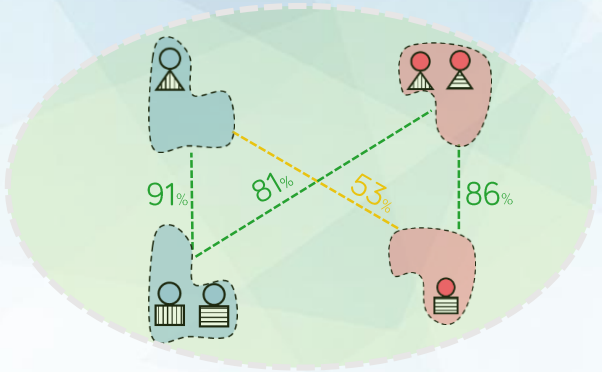
some items



Similarity over a set of items: sim is depicted by an average (or a weighted average) of $simobj$ over a given set of items

COMPARING MODELS

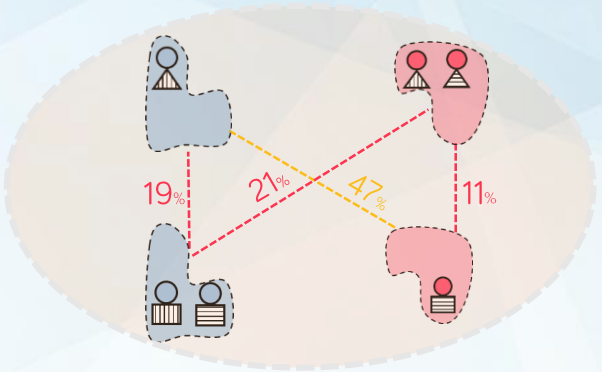
Global pairwise behavior

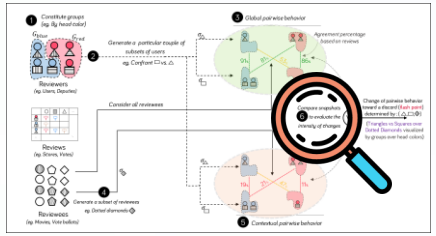


Comparison between models

Use of quality measure

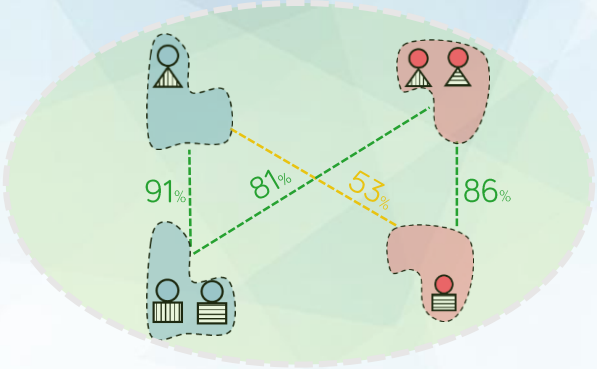
Contextual pairwise behavior





COMPARING MODELS

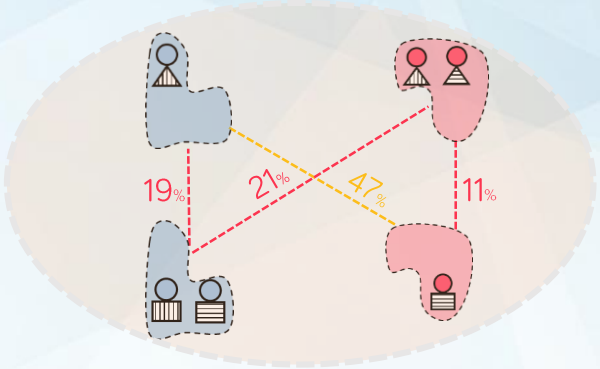
Global pairwise behavior

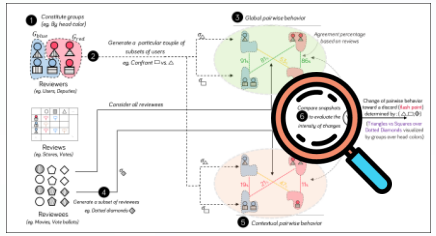


Comparison between models

Use of quality measure

Contextual pairwise behavior



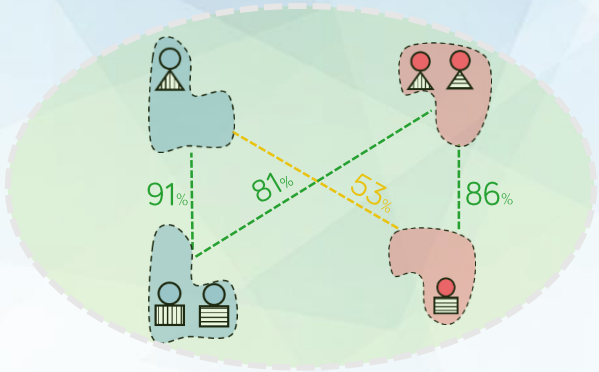


$\varphi_{consent}$: to uncover contexts that leads to more agreement

$$\varphi_{consent} = \frac{\sum_{(i,j)\in\gamma_L(U_{g'})\times\gamma_L(U_{g''})} \max(sim(E_{context},i,j) - sim(E,i,j),0)}{|\gamma_L(U_{g'})|\times|\gamma_L(U_{g''})|}$$

COMPARING MODELS

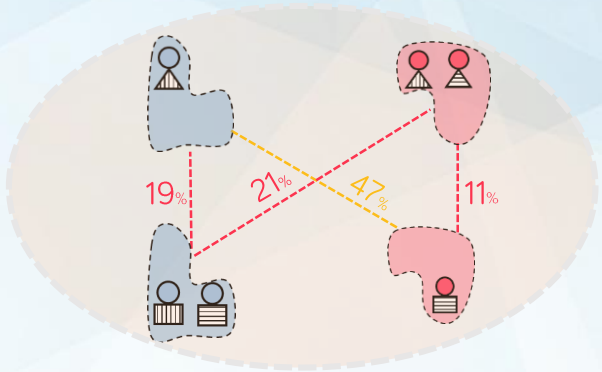
Global pairwise behavior



Comparison between models

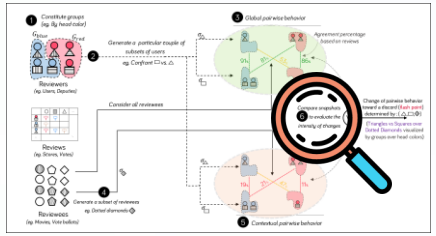
Use of quality measure

Contextual pairwise behavior



$\varphi_{consent}$: to uncover contexts that leads to more agreement

$\varphi_{dissent}$: to uncover contexts that leads to more disagreement



$$\varphi_{dissent} = \frac{\sum_{(i,j)\in\gamma_L(U_{g'})\times\gamma_L(U_{g''})} \max(sim(E,i,j) - sim(E_{context},i,j),0)}{|\gamma_L(U_{g'})|\times|\gamma_L(U_{g''})|}$$

SUMMARY

- DSC (Discovering similarities change) Framework
- **DSC algorithm**
- Use cases

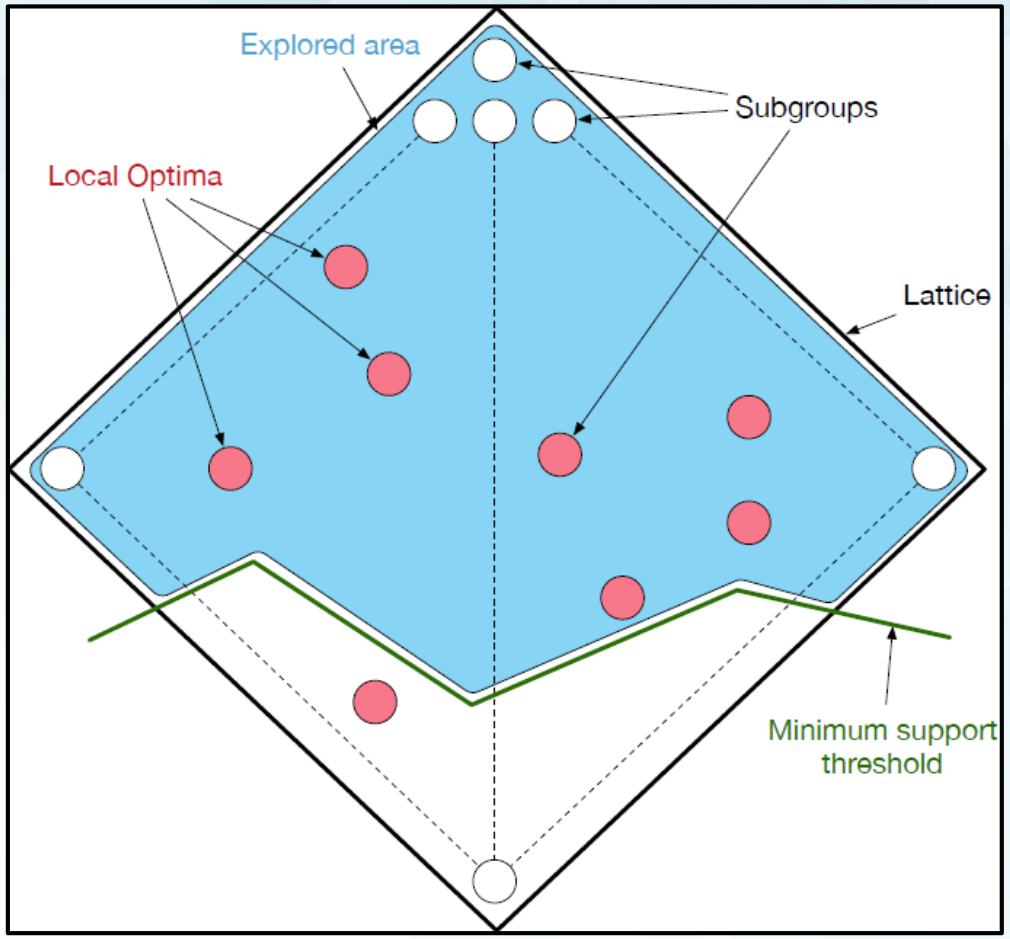
ALGORITHM

Enumerating candidate subgroups:

In order to traverse the search space efficiently while having the guarantees to obtain the real top k patterns we need to:

- Efficiency of enumeration
  (Enumerate each subgroup only once)
- Effectiveness of pruning techniques: Prune unpromising sub-search space ASAP.

Enumerate each subgroup at most once

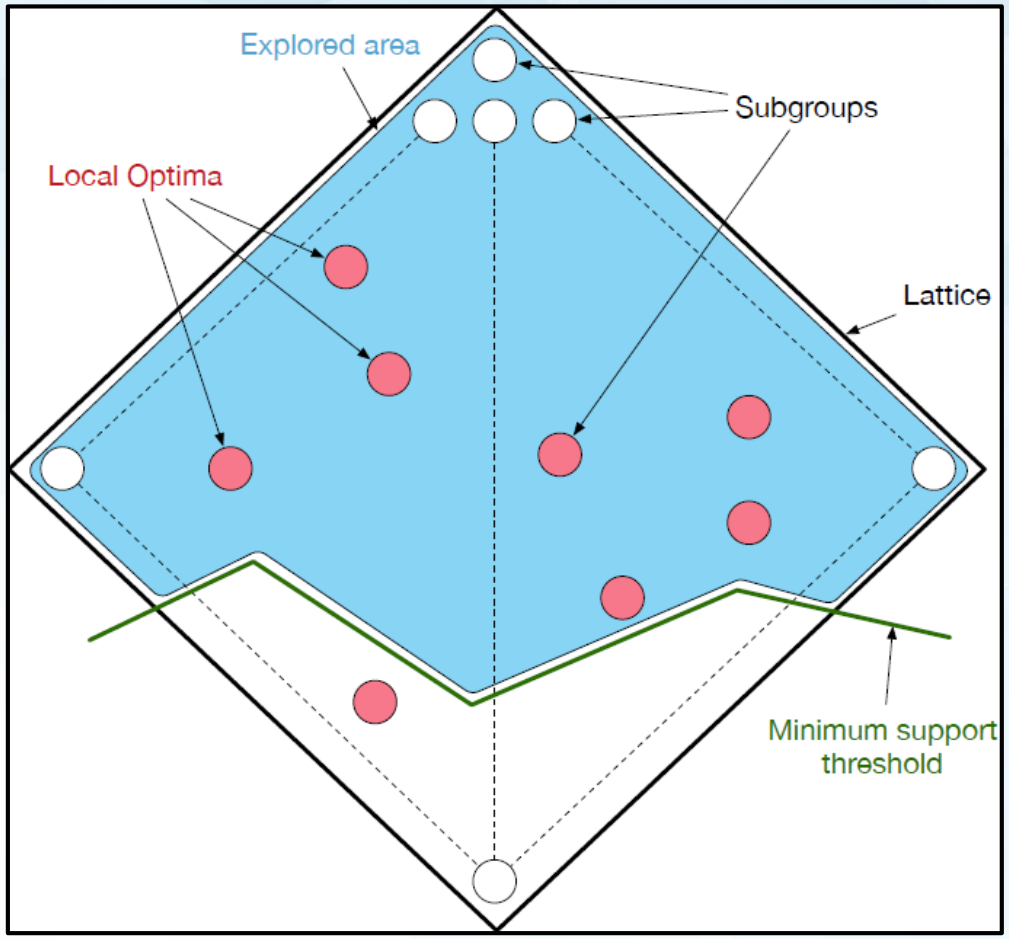


ALGORITHM

Enumerating candidate subgroups:

In order to traverse the search space efficiently while having the guarantees to obtain the real top k patterns we need to:

- Avoid redundancy in Enumeration: take benefit of closed descriptions and impose a canonical order between descriptions.
- Prune unpromising sub-search space ASAP: define upper bounds and propose a branch and bound algorithm



ALGORITHM

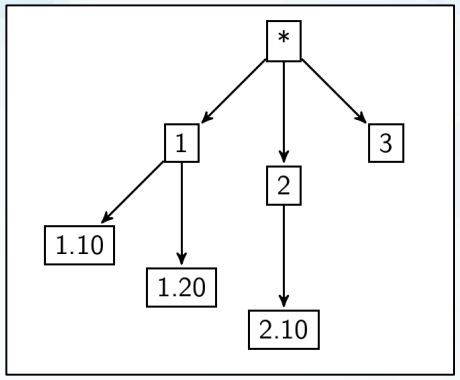
a) Closed descriptions: We defined a closure operator over a complex description (different types of attributes) relying on closure over each attribute (Boolean, Categorical, numerical*)

*M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. Information Sciences,181(10):1989{2001, 2011.

ALGORITHM

a) Closed descriptions: We defined a closure operator over a complex description (different types of attributes) relying on closure over each attribute (Boolean, Categorical, numerical*, HMT)

- New pattern domain: HMT (Hierarchical Multi-tag attribute) relying on pattern structures†



	$tags$
g_1	$\{1.20, 2.10\}$
g_2	$\{1, 3\}$
g_3	$\{1.10, 2.10, 3\}$
g_4	$\{2.10\}$
g_5	$\{1.20\}$

*M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. Information Sciences,181(10):1989{2001, 2011.

† Ganter, B., & Kuznetsov, S. (2001). Pattern structures and their projections. Conceptual Structures: Broadening the Base, 129-142

ALGORITHM – HANDLING HMT

New pattern domain: HMT (Hierarchical Multi-tag attribute) relying on pattern structures

Items (Ballots) - E	
IdBallot	*Theme (HMT)*
001	1.10 Justice 2.10 Europe coop
002	1.10 Justice 2.20 Europe inter.
003	2.30 Consumer Protection 3.10 Agriculture
004	3.10 Agriculture

European Parliament Datasets: 378 Tags

Items (Places) - E	
IdPlace	*Theme (HMT)*
130	07.24.03. Session Photography 14.38. IT Services & Computer Repair
2012	09.39 Grocery 22.19 Drugstores
4231	21.24 Breakfast & Brunch 21.26 Buffets
31425	09.03 Bagels 09.15 Coffee & Tea

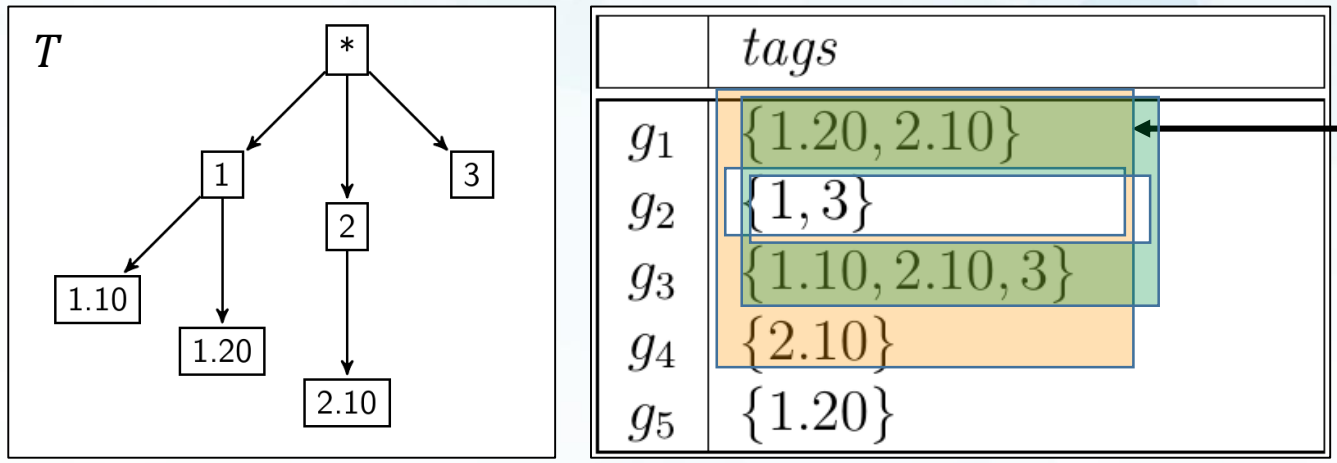
YELP: 1509 Tags

ALGORITHM – HANDLING HMT

HMT (Hierarchical Multi-tag attribute) Descriptions (Conjunction of tags)

$d = \{2\}$ | $d' = \{g1, g3, g4\}$

$d = \{1,2\}$ | $d' = \{g1, g3\}$



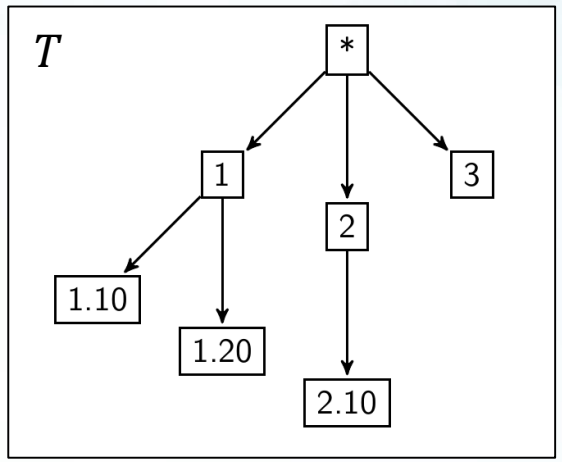
ALGORITHM – HANDLING HMT

HMT (Hierarchical Multi-tag attribute) Descriptions (Conjunction of tags)

$d = \{2\}$ | $d' = \{g1, g3, g4\}$

$d = \{1,2\}$ | $d' = \{g1, g3\}$

Pattern Ordinal Scaling



	tags
g_1	$\{1.20, 2.10\}$
g_2	$\{1, 3\}$
g_3	$\{1.10, 2.10, 3\}$
g_4	$\{2.10\}$
g_5	$\{1.20\}$

Take into account the taxonomy ***T***

	$*$	1	1.10	1.20	2	2.10	3
g_1	×	×		×	×	×	
g_2	×	×					×
g_3	×	×	×		×		×
g_4	×				×	×	
g_5	×	×		×			

Ignore the taxonomy ***T***

ALGORITHM – HANDLING HMT

Pattern Structure† $(\mathbf{G}, (\mathbf{D}, \sqcap), \boldsymbol{\delta})$

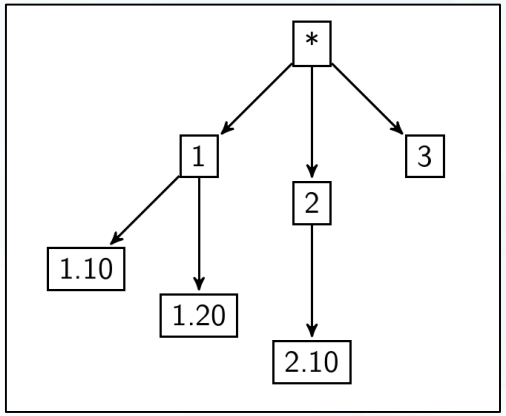
$\mathbf{G}$: a set of objects

$\mathbf{D}$: Domain of descriptions (*Conjunction of tags*)

$\sqcap$: Infimum (Similarity Operator), (D, $\sqcap$) forms a semi-lattice, the partial order denoted $\sqsubseteq$. We have $\boldsymbol{c} \sqsubseteq \boldsymbol{d} \iff \boldsymbol{c} \sqcap \boldsymbol{d} = \boldsymbol{c}$

$\boldsymbol{\delta}$: A mapping function $\boldsymbol{\delta}: \boldsymbol{G} \rightarrow \boldsymbol{D}$

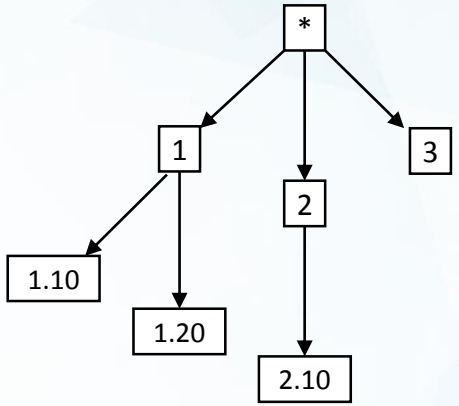
	tags
g_1	$\{1.20, 2.10\}$
g_2	$\{1, 3\}$
g_3	$\{1.10, 2.10, 3\}$
g_4	$\{2.10\}$
g_5	$\{1.20\}$



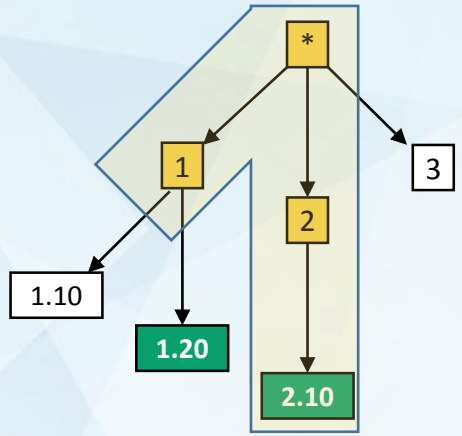
† Ganter, B., & Kuznetsov, S. (2001). Pattern structures and their projections. Conceptual Structures: Broadening the Base, 129-142

ALGORITHM – HANDLING HMT

Defining the infimum (similarity) operator for $(\mathbf{D}, \sqcap)$

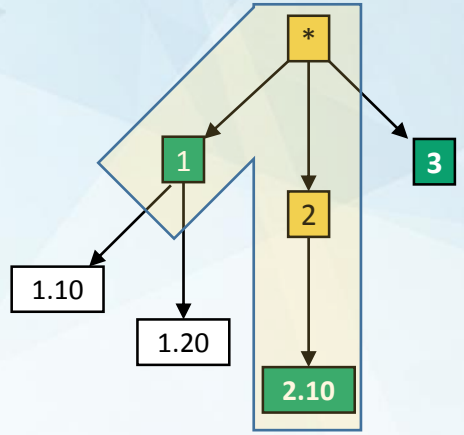


Whole hierarchy

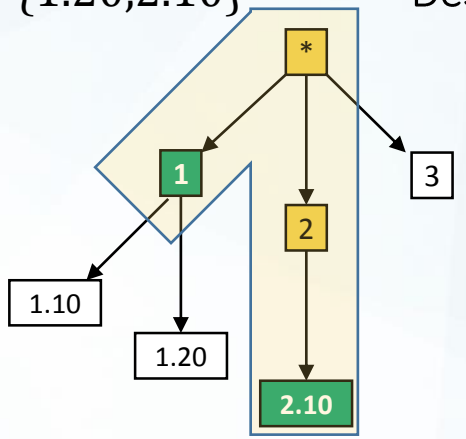


Description $c = \{1.20, 2.10\}$

$\sqcap$



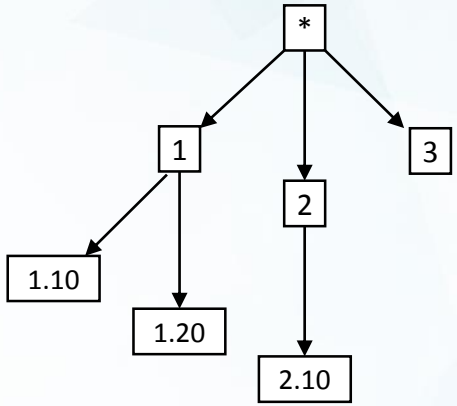
Description $d = \{1, 2.10, 3\}$



$c \sqcap d = \{1, 2.10\}$

ALGORITHM – HANDLING HMT

Defining the infimum (similarity) operator for $(\mathbf{D}, \sqcap)$

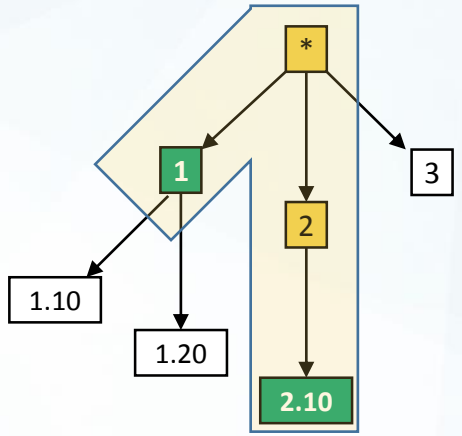


Whole hierarchy

Infimum $\sqcap_{HMT}$ definition

Given c, d two descriptions from D:

- $c \sqcap_{HMT} d = \max(\cup_{t \in c}\uparrow t \cap \cup_{u \in d}\uparrow u)$
- $max(s) = \{t \in s \mid (\downarrow t - \{t\}) \cap s = \emptyset\}$

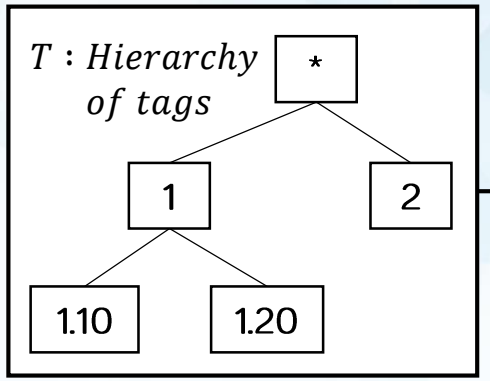


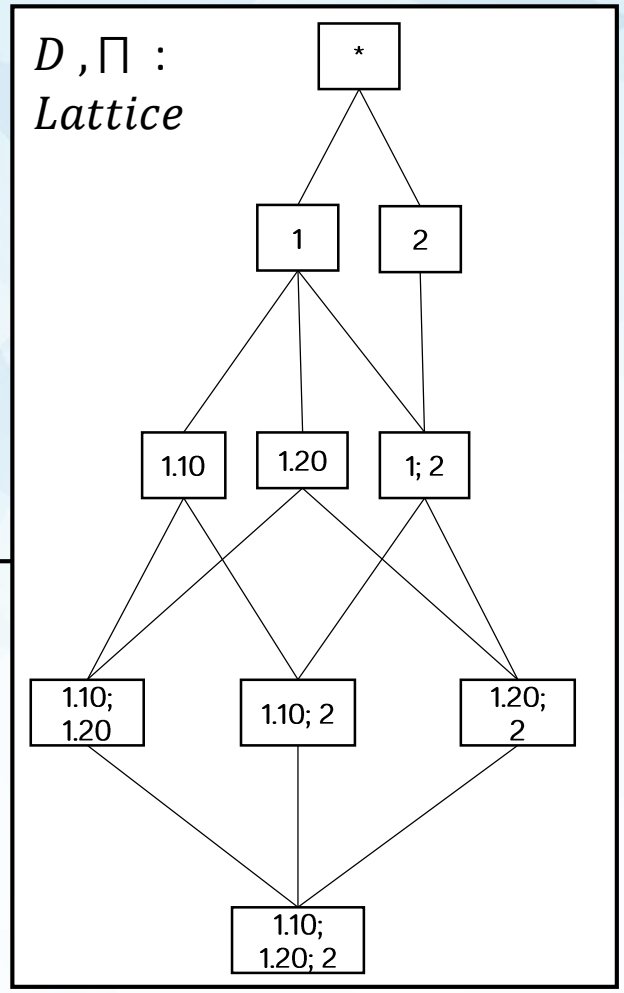
$c \sqcap d = \{1, 2.10\}$

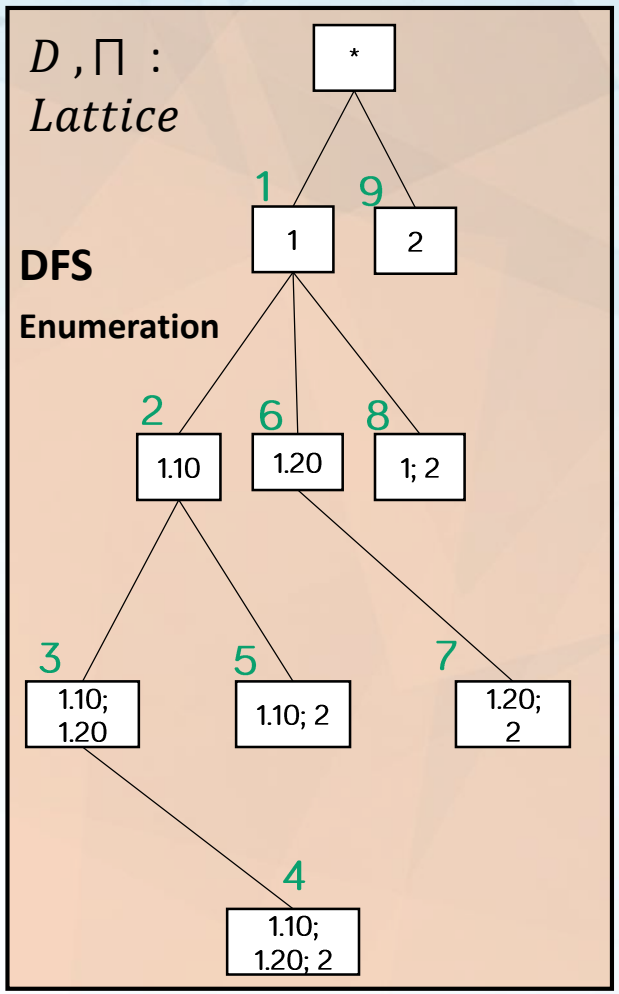
Allows us to define straightforwardly a closure operator over HMT descriptions

ALGORITHM – HANDLING HMT

Enumerating pattern using hierarchical Structure of tags $\boldsymbol{T}$.







Perf example : CbO

(155 tags, max_tags_by_obj=3, 100 objects, depth=3, 5-ary)

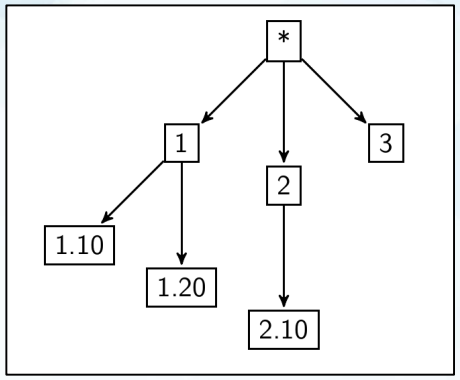
Itemset (Scaling): 12737 visited for 243 closed

HMT: 1823 visited for 243 closed

ALGORITHM

a) Closed descriptions: We defined a closure operator over a complex description (different types of attributes) relying on closure over each attribute (Boolean, Categorical, numerical*, HMT)

- New pattern domain: HMT (Hierarchical Multi-tag attribute) relying on pattern structures†



	$tags$
g_1	$\{1.20, 2.10\}$
g_2	$\{1, 3\}$
g_3	$\{1.10, 2.10, 3\}$
g_4	$\{2.10\}$
g_5	$\{1.20\}$

*M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. Information Sciences,181(10):1989{2001, 2011.

† Ganter, B., & Kuznetsov, S. (2001). Pattern structures and their projections. Conceptual Structures: Broadening the Base, 129-142

ALGORITHM

a) Closed descriptions: We defined a closure operator over a complex description (different types of attributes) relying on closure over each attribute (Boolean, Categorical, numerical*, HMT)

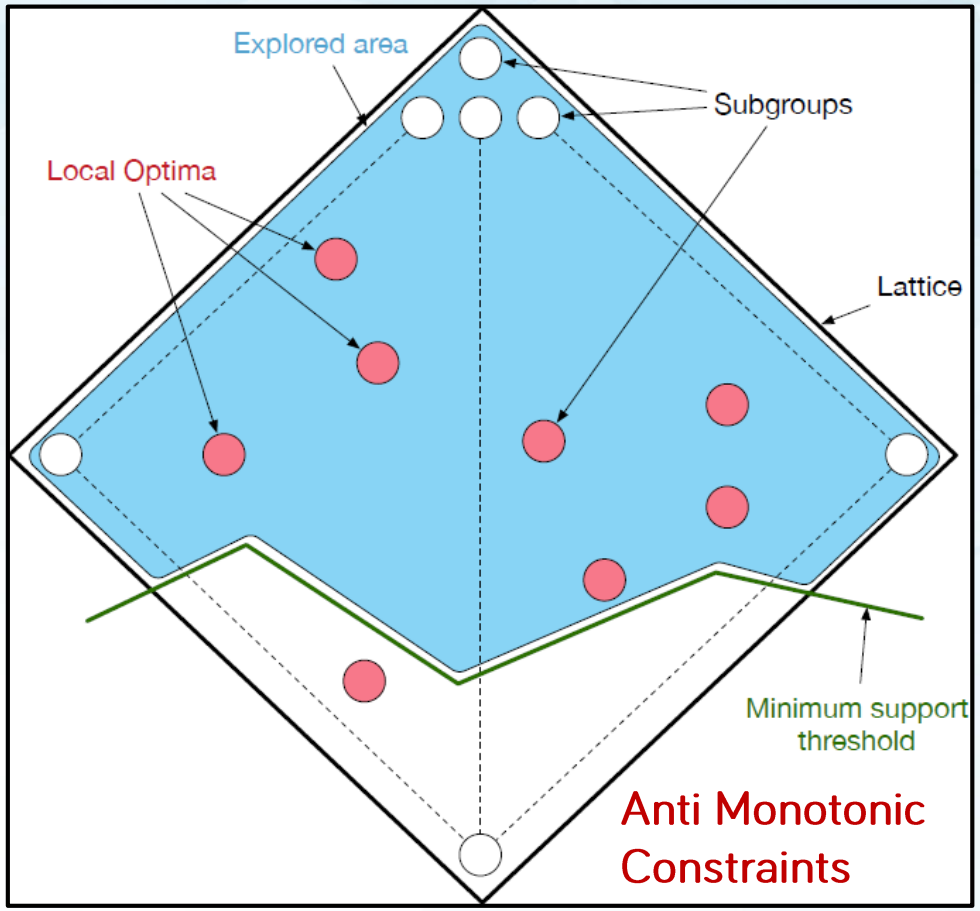
- New pattern domain: HMT (Hierarchical Multi-tag attribute) relying on pattern structures†
- Adapt CbO (close-by-one**) algorithm to enumerate candidate subgroups in a DFS fashion (Without redundancy)

** S. O. Kuznetsov. Learning of simple conceptual graphs from positive and negative examples. In PKDD, pages 384{391. Springer, 1999.

*M. Kaytoue, S. O. Kuznetsov, A. Napoli, and S. Duplessis. Mining gene expression data with pattern structures in formal concept analysis. Information Sciences,181(10):1989{2001, 2011.

† Ganter, B., & Kuznetsov, S. (2001). Pattern structures and their projections. Conceptual Structures: Broadening the Base, 129-142

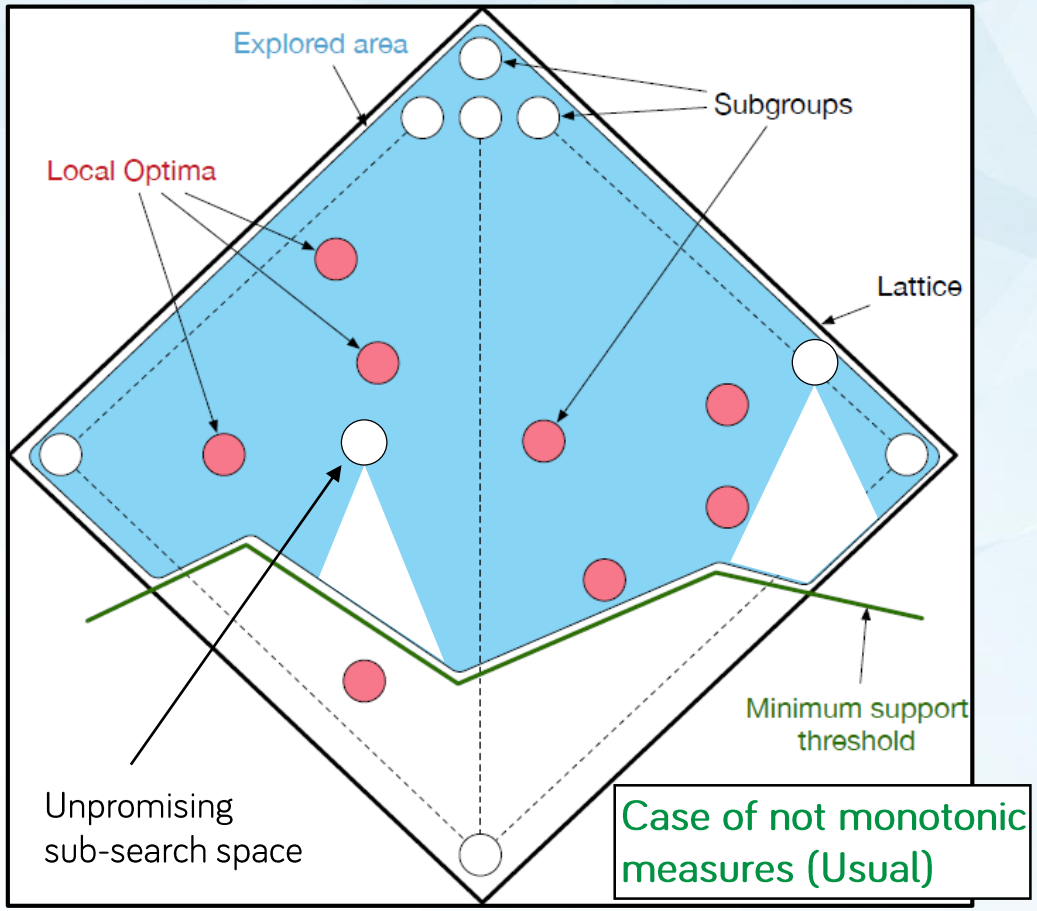
b) Branch and bound algorithm: definition of upper bounds on both quality measures for a more efficient enumeration.



ALGORITHM

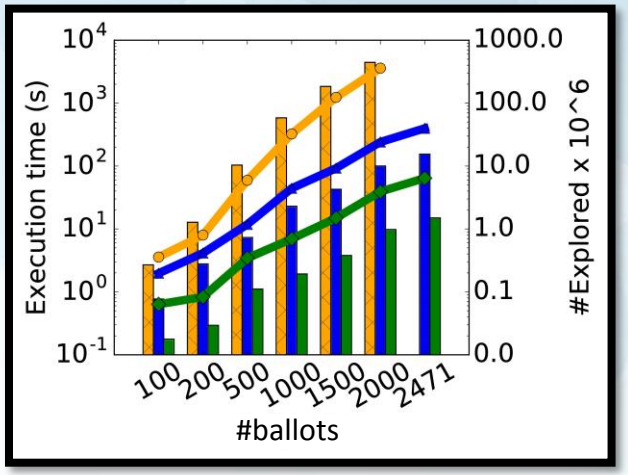
b) Branch and bound algorithm: definition of upper bounds on both quality measures for a more efficient enumeration.

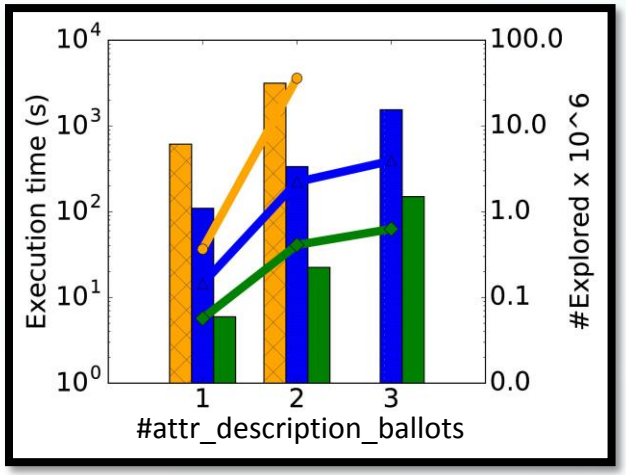
- Evaluate the upper bound $ub(p)$ of each **generated subgroup p** (by adapted CbO)
- If $\boldsymbol{ub(p) < \sigma_{\varphi}}$ then backtrack and do not consider all the sub search space of p i.e. $\{p' \in D \mid p \sqsubset p'\}$
- We defined two upper bounds of the quality measure:
  - UB1 less tight but not costly
  - UB2 tighter but costly



DSC PERFORMANCE

DECADE
Belfodil, Lamarre, Cazalens & Plantevit





European parliament datasets

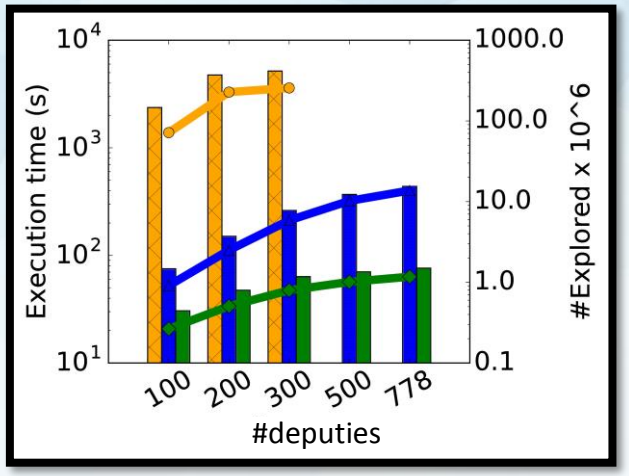
784 Deputies

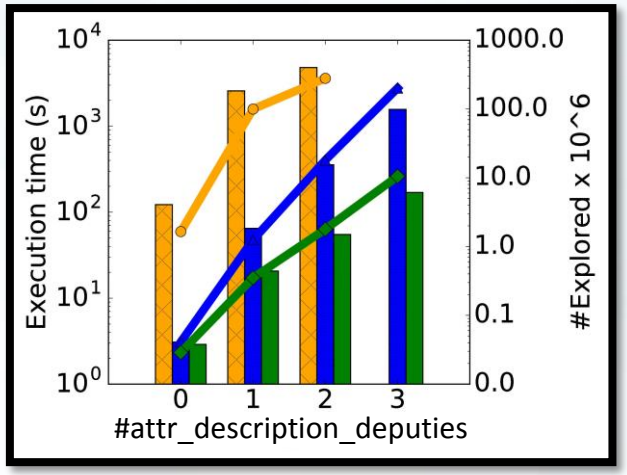
3 Categorical

2471 ballots

1 HMT + 1 Numerical + 1 Categorical

1 600 000 votes





Baseline

Closed

Branch & bound + Closed DSC (UB1)

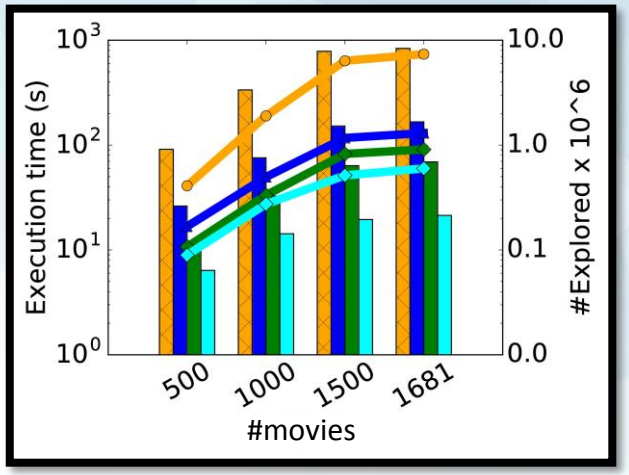
Execution time

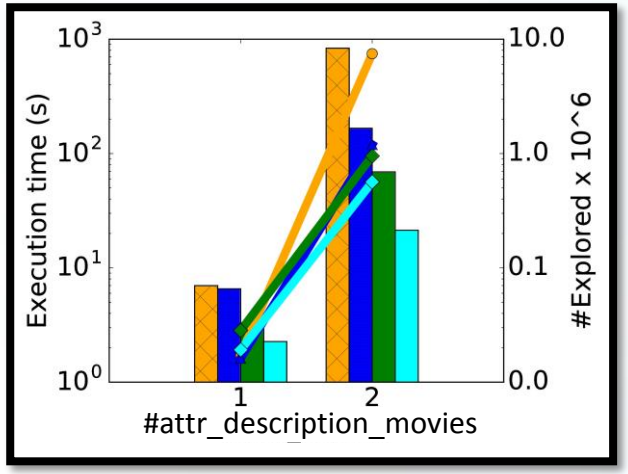
#Explored descriptions

CONTENTCHECK

DSC PERFORMANCE

DECADE
Belfodil, Lamarre, Cazalens & Plantevit





Movielens dataset

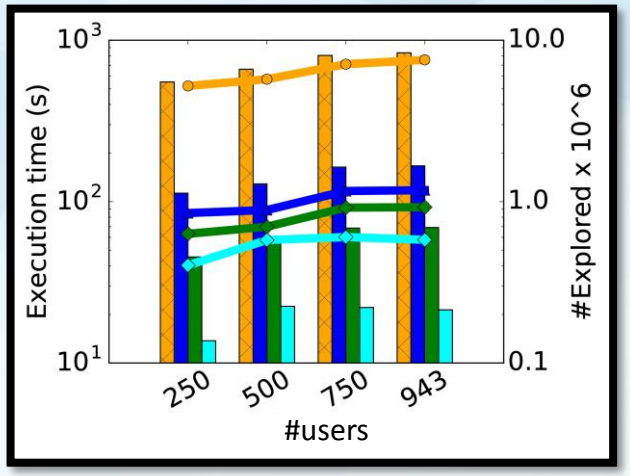
943 Users

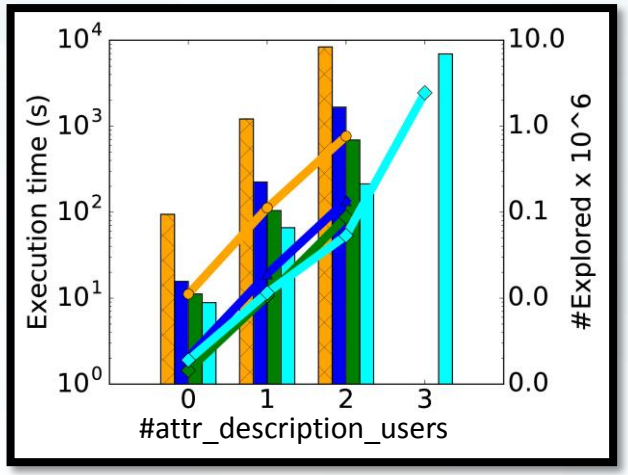
3 Categorical

1681 Movies

1 HMT + 1 Numerical

100 000 ratings





Baseline

Closed

Branch & bound + Closed DSC (UB1)

Branch & bound + Closed DSC (UB2)

Execution time

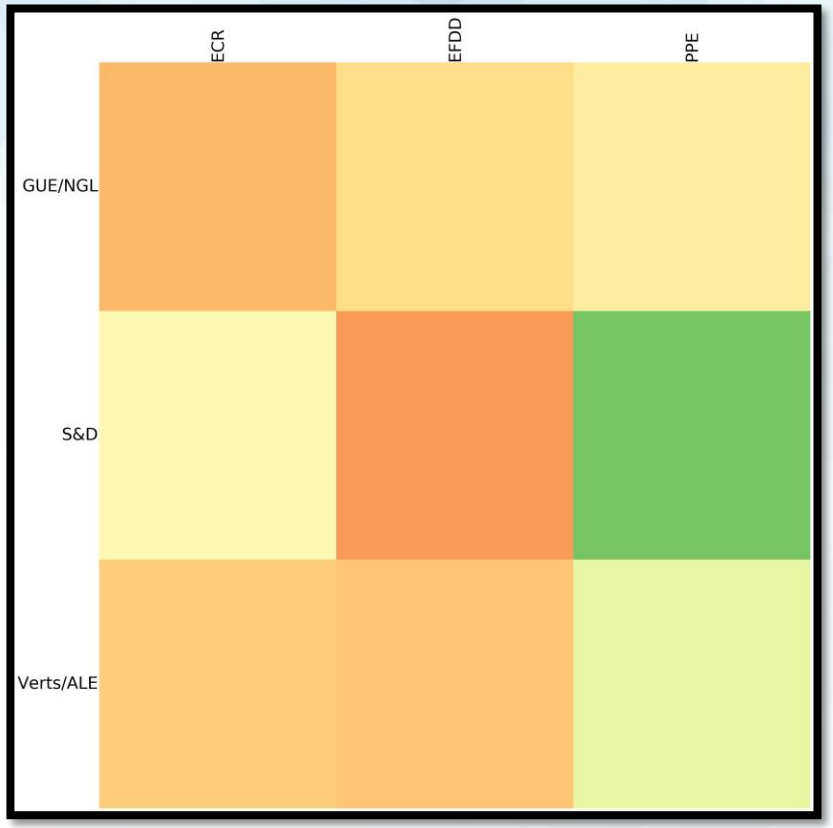
#Explored descriptions

SUMMARY

- DSC (Discovering similarities change) Framework
- DSC algorithm
- **Use cases**

EXAMPLE: EUROPEAN PARLIAMENT DATASET

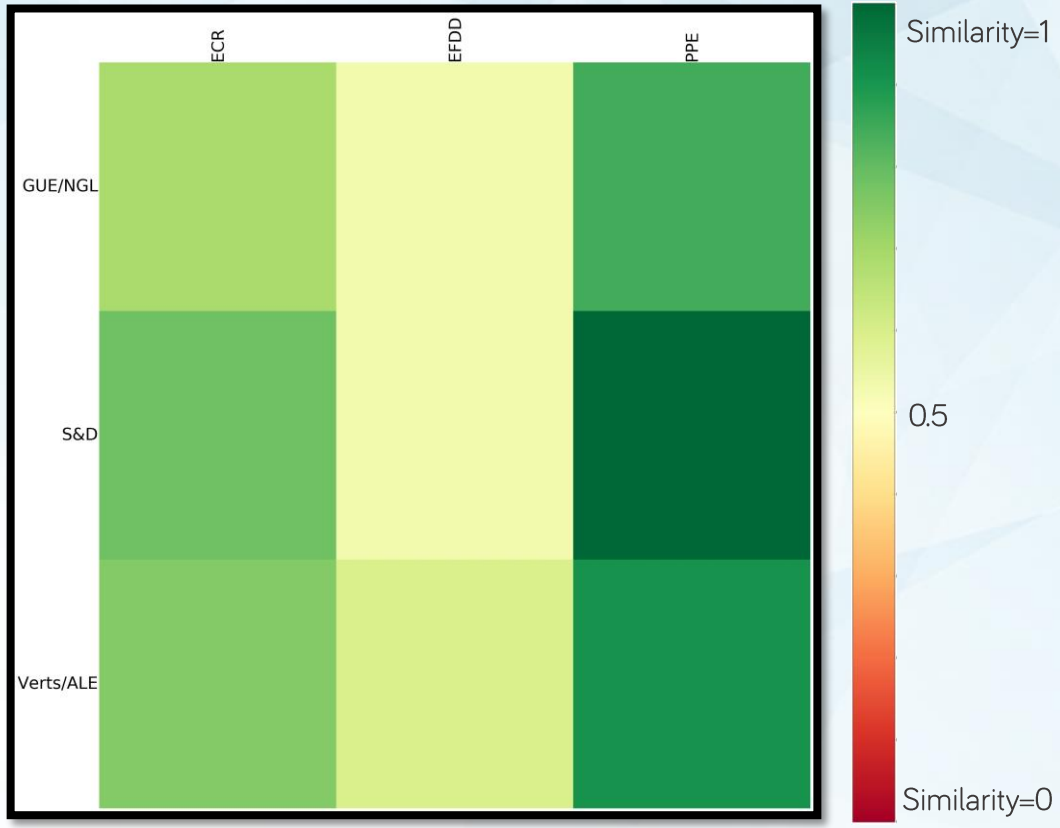
Usual pairwise behavior



Toward consent between European political groups
The pattern:

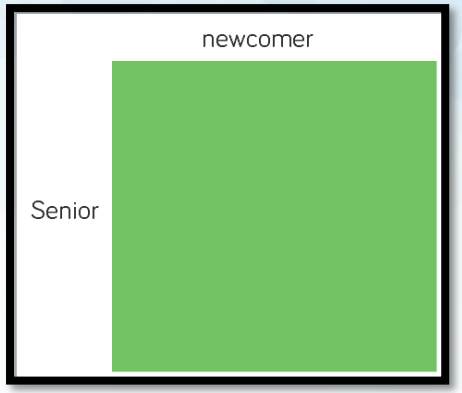
[7.40 European judicial conventions during Feb. – Nov. 2015 ,left wing ,right wing]

Contextual pairwise behavior



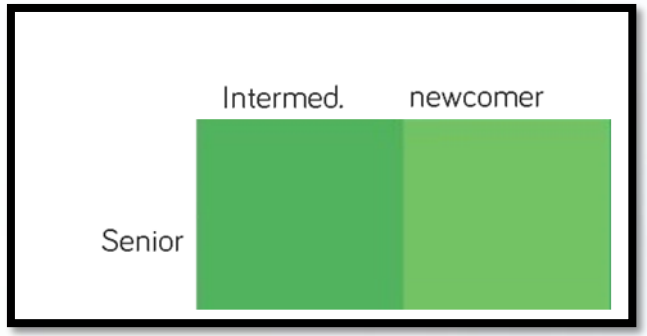
EXAMPLE: YELP DATASET

Usual pairwise behavior



Toward Dissent between Yelp Users for the context:

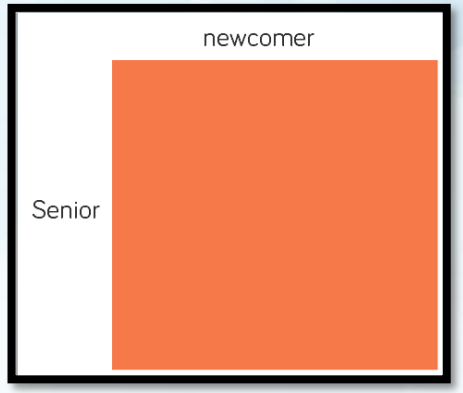
[Professional Services, Shopping, In Oklahoma, Senior, Newcomer]

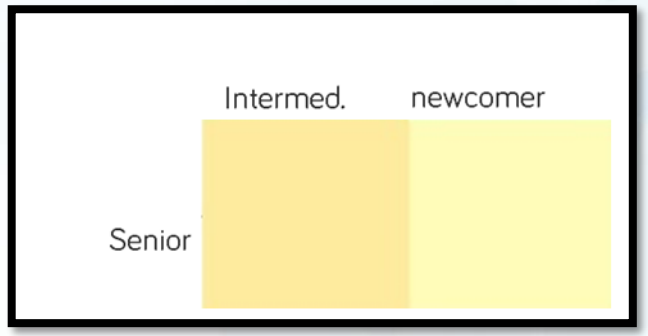


Toward Dissent between Yelp Users for the context:

[Medical Center, Doctors, In Wisconsin, {Senior}, {Interm., Newcomer},]

Contextual pairwise behavior





SUMMARY

Conclusion & Perspectives

Conclusion

- Definition of a novel problem :
  discovering exceptional pairwise behavior
- Implementation of a branch and bound enumeration algorithm (DSC)
- Experimentation over real-world datasets (voting and rating)

CONCLUSION AND PERSPECTIVES

Conclusion

- Definition of a novel problem : discovering exceptional pairwise behavior
- Implementation of a branch and bound enumeration algorithm (DSC)
- Experimentation over real-world datasets (voting and rating)

Perspectives & current work

- Providing adapted instant mining and interactive mining algorithms.
- Studying the behavior of groups of individuals (deputies) through time.
- ContentCheck: Offering a set of tools for journalists (Le Monde) of lead finding/fact checking
- Extend HMT to handle datasets where implications between items are known.

THANKS

FOR YOUR TIME

QUESTIONS

Contact : adnene.belfodil@insa-lyon.fr

Materials: https://github.com/Adnene93/DiscoveringSimilarityChanges

Feel free to ask any question you have :-)

BIBLIOGRAPHY

- W. Duivesteijn, A. J. Feelders, and A. Knobbe. Exceptional model mining. Data Mining and Knowledge Discovery, 30(1):47{98, 2016.
- B. Ganter and S. O. Kuznetsov. Pattern structures and their projections. ICCS,2001.
- C. R. de Sà, W. Duivesteijn, C. Soares, and A. Knobbe. Exceptional preferences mining. In DS, pages 3{18. Springer, 2016.
- G. Bosc, J. Golebiowski, M. Bensa, C. Robardet, M. Plantevit, J. Boulicaut, and M. Kaytoue. Local subgroup discovery for eliciting and understanding new structure-odor relationships. In DS, pages 19{34, 2016.
- S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. Lakshmanan, and R. H. Zamar. Exploring rated datasets with rating maps. In WWW, 2017.
- M. Das, S. Amer-Yahia, G. Das, and C. Yu. Mri: Meaningful interpretations of collaborative ratings. PVLDB

THANKS FOR YOUR TIME

? QUESTIONS

Contact : adnene.belfodil@insa-lyon.fr

Materials: https://github.com/Adnene93/DiscoveringSimilarityChanges

Feel free to ask any question you have :-)