JOCLAD 2021

UNIVERSIDADE
BEIRA INTERIOR

Inter-Organizational Networks of the EuroGroups Register

A Supervised Clustering Algorithm for Network Data

Bárbara Monteiro Santos, barbara.monteiro@ine.pt
Pedro Campos, pedro.campos@ine.pt

Statistics Portugal

CLAD
Associação Portuguesa de Classificação e Análise de Dados

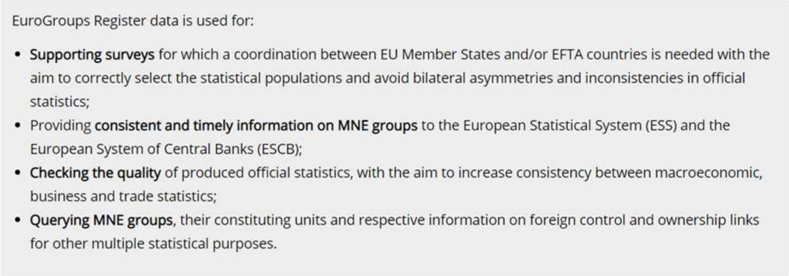
Instituto Nacional de Estatística
Statistics Portugal

TABLE OF CONTENTS

- Introduction and Motivation
- Literature Review: Main Lines
- SUWAN Algorithm
- EGR Network
- Implementation
- Conclusions and Limitations
- Bibliography

INTRODUCTION AND MOTIVATION

Together with the national statistical business registers (NSBRs), the EuroGroups Register is part of the **European Framework of business registers**.
The EuroGroups Register exchange confidential data on legal units, enterprises and enterprise groups containing information on **multinational enterprises (MNE)** groups operating in Europe.



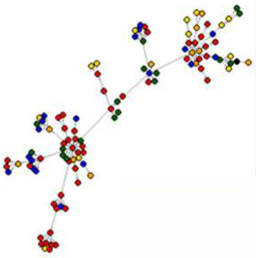
EuroGroups Register data is used for:

- **Supporting surveys** for which a coordination between EU Member States and/or EFTA countries is needed with the aim to correctly select the statistical populations and avoid bilateral asymmetries and inconsistencies in official statistics;
- Providing **consistent and timely information on MNE groups** to the European Statistical System (ESS) and the European System of Central Banks (ESCB);
- **Checking the quality** of produced official statistics, with the aim to increase consistency between macroeconomic, business and trade statistics;
- **Querying MNE groups**, their constituting units and respective information on foreign control and ownership links for other multiple statistical purposes.

Source: Eurostat

INTRODUCTION AND MOTIVATION

- **Multinational enterprise (MNE) groups can be seen as networks**



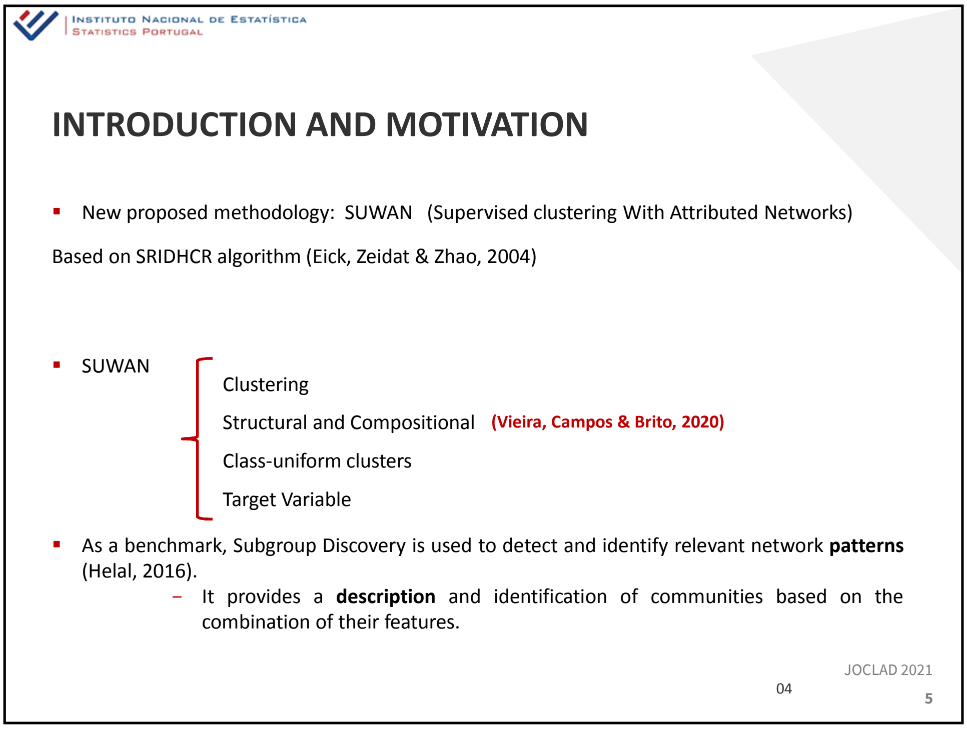
- **Goal: Distinguishing characteristics of social networks: propensity for displaying community structure**

Community Detection

- Find cohesive subgraphs of nodes
- **Structural Aspects** (Harenberg et al., 2014)

Clustering

- Divide a set of objects into homogeneous groups
- **Compositional Characteristics**

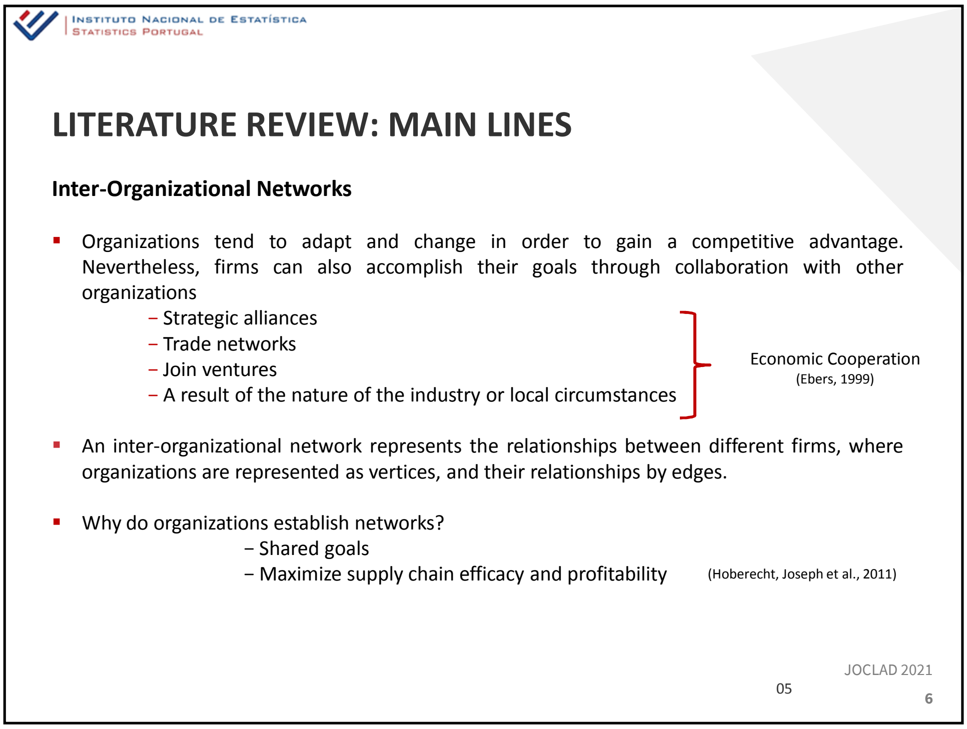


INTRODUCTION AND MOTIVATION

- New proposed methodology: SUWAN (Supervised clustering With Attributed Networks)

Based on SRIDHCR algorithm (Eick, Zeidat & Zhao, 2004)

- SUWAN
  - Clustering
  - Structural and Compositional **(Vieira, Campos & Brito, 2020)**
  - Class-uniform clusters
  - Target Variable
- As a benchmark, Subgroup Discovery is used to detect and identify relevant network **patterns** (Helal, 2016).
  - It provides a **description** and identification of communities based on the combination of their features.



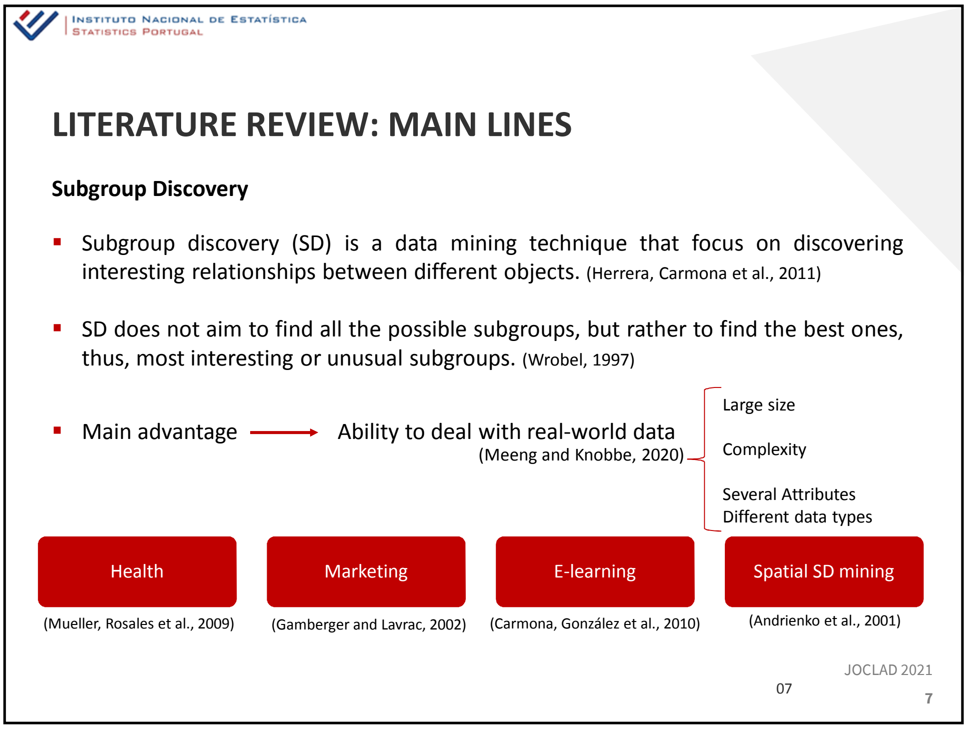
LITERATURE REVIEW: MAIN LINES

Inter-Organizational Networks

- Organizations tend to adapt and change in order to gain a competitive advantage. Nevertheless, firms can also accomplish their goals through collaboration with other organizations
  - Strategic alliances
  - Trade networks
  - Join ventures
  - A result of the nature of the industry or local circumstances

Economic Cooperation
(Ebers, 1999)

- An inter-organizational network represents the relationships between different firms, where organizations are represented as vertices, and their relationships by edges.
- Why do organizations establish networks?
  - Shared goals
  - Maximize supply chain efficacy and profitability (Hoberecht, Joseph et al., 2011)



Instituto Nacional de Estatística
Statistics Portugal

LITERATURE REVIEW: MAIN LINES

Subgroup Discovery

- Subgroup discovery (SD) is a data mining technique that focus on discovering interesting relationships between different objects. (Herrera, Carmona et al., 2011)
- SD does not aim to find all the possible subgroups, but rather to find the best ones, thus, most interesting or unusual subgroups. (Wrobel, 1997)
- Main advantage → Ability to deal with real-world data (Meeng and Knobbe, 2020)

Large size

Complexity

Several Attributes
Different data types

Health (Mueller, Rosales et al., 2009)

Marketing (Gamberger and Lavrac, 2002)

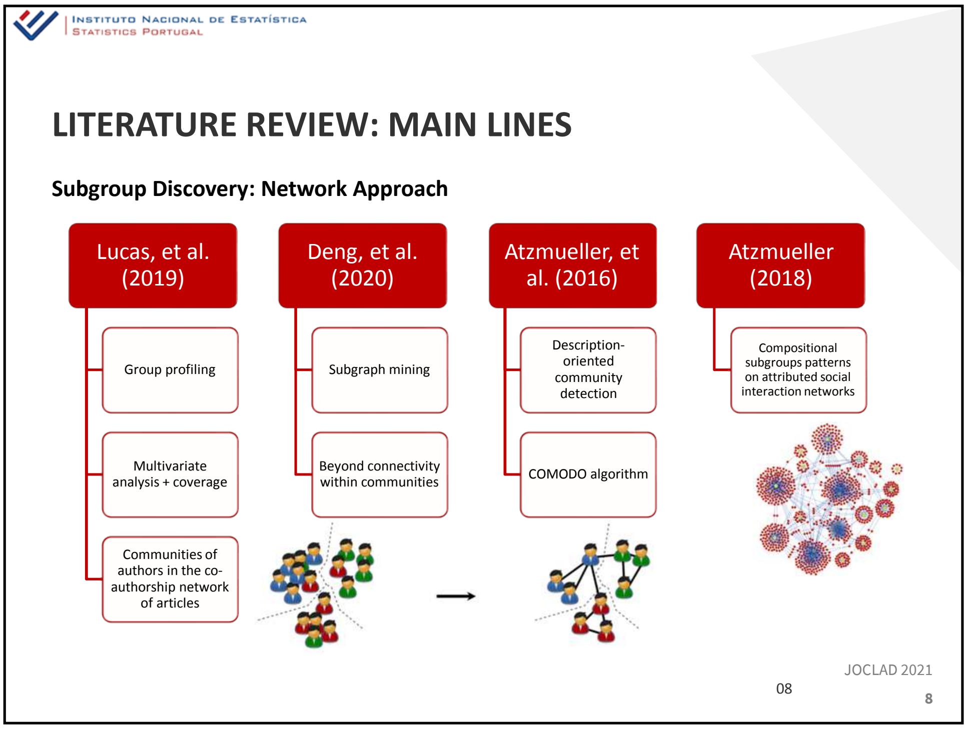
E-learning (Carmona, González et al., 2010)

Spatial SD mining (Andrienko et al., 2001)

JOCLAD 2021

07

7



Instituto Nacional de Estatística
Statistics Portugal

LITERATURE REVIEW: MAIN LINES

Subgroup Discovery: Network Approach

Lucas, et al. (2019)
- Group profiling
- Multivariate analysis + coverage
- Communities of authors in the co-authorship network of articles

Deng, et al. (2020)
- Subgraph mining
- Beyond connectivity within communities

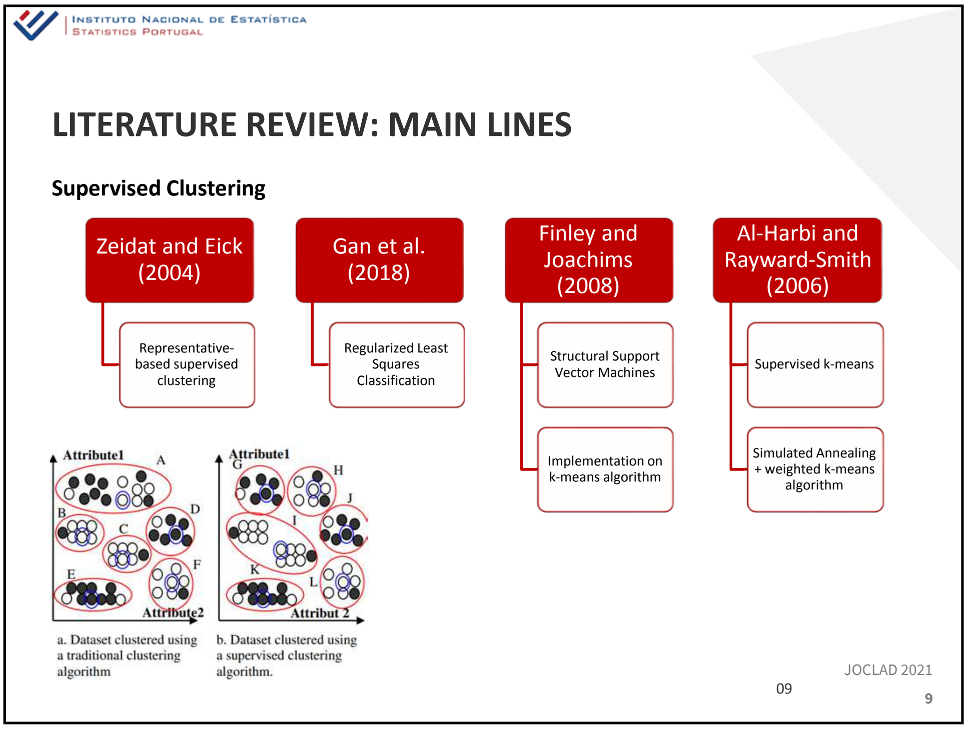
Atzmueller, et al. (2016)
- Description-oriented community detection
- COMODO algorithm

Atzmueller (2018)
- Compositional subgroups patterns on attributed social interaction networks

JOCLAD 2021

08

8



Instituto Nacional de Estatística
Statistics Portugal
Attribute1
A
B
C
D
E
F
Attribute2
Attribute1
G
H
I
J
K
L
Attribut 2

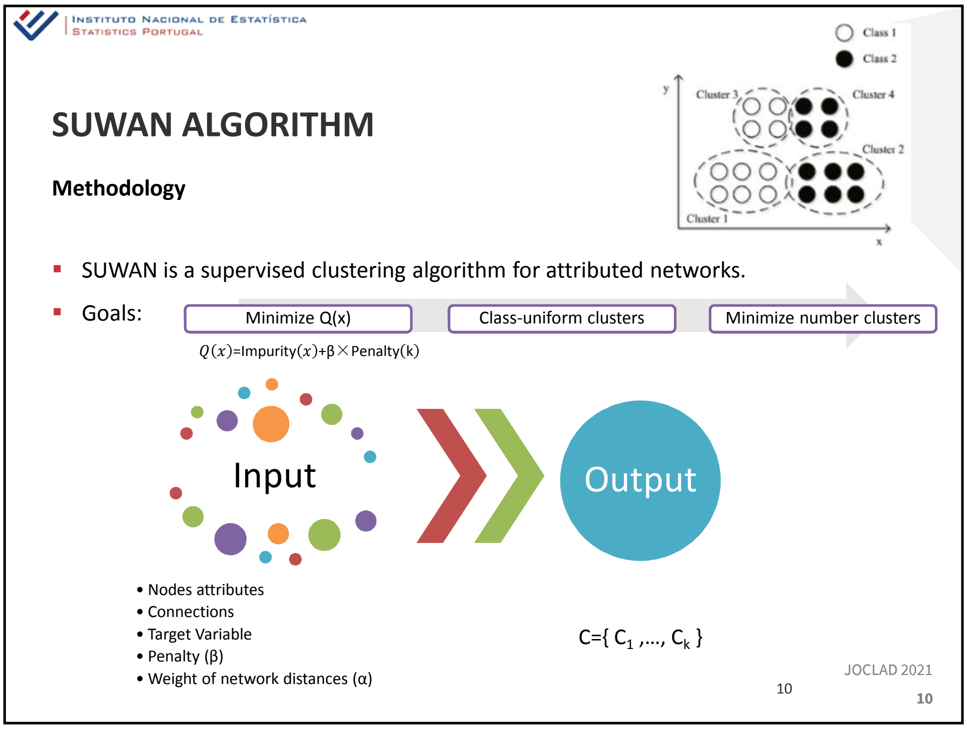
LITERATURE REVIEW: MAIN LINES

Supervised Clustering

- Zeidat and Eick (2004)
  - Representative-based supervised clustering
- Gan et al. (2018)
  - Regularized Least Squares Classification
- Finley and Joachims (2008)
  - Structural Support Vector Machines
  - Implementation on k-means algorithm
- Al-Harbi and Rayward-Smith (2006)
  - Supervised k-means
  - Simulated Annealing + weighted k-means algorithm

a. Dataset clustered using a traditional clustering algorithm

b. Dataset clustered using a supervised clustering algorithm.



Instituto Nacional de Estatística
Statistics Portugal
Class 1
Class 2
y
Cluster 3
Cluster 4
Cluster 2
Cluster 1
x

SUWAN ALGORITHM

Methodology

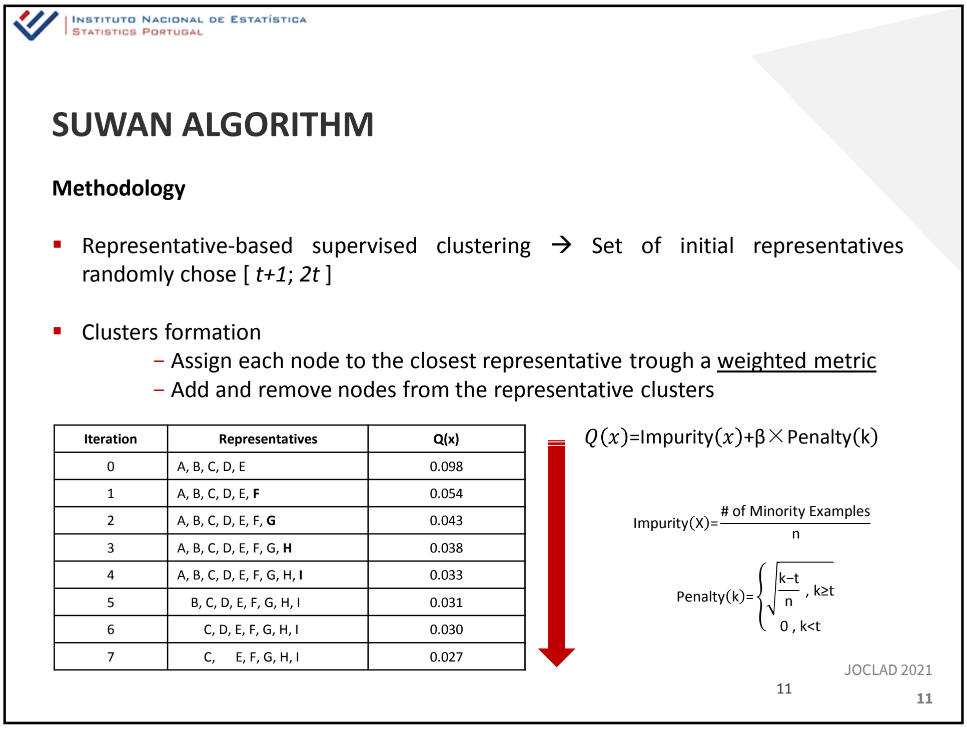
- SUWAN is a supervised clustering algorithm for attributed networks.
- Goals: Minimize Q(x) → Class-uniform clusters → Minimize number clusters

$Q(x)$=Impurity(x)+β×Penalty(k)

Input → Output

- Nodes attributes
- Connections
- Target Variable
- Penalty (β)
- Weight of network distances (α)

C={ C_1 ,…, C_k }



Instituto Nacional de Estatística
Statistics Portugal

SUWAN ALGORITHM

Methodology

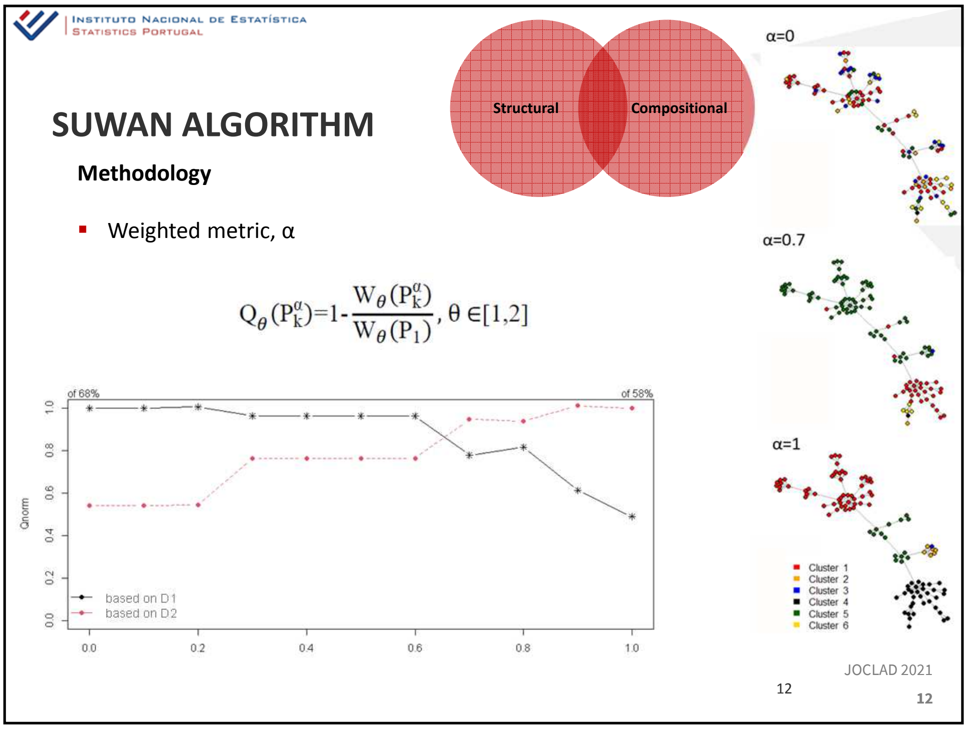
- Representative-based supervised clustering → Set of initial representatives randomly chose [*t+1*; *2t*]
- Clusters formation
  - Assign each node to the closest representative trough a weighted metric
  - Add and remove nodes from the representative clusters

Iteration	Representatives	Q(x)
0	A, B, C, D, E	0.098
1	A, B, C, D, E, **F**	0.054
2	A, B, C, D, E, F, **G**	0.043
3	A, B, C, D, E, F, G, **H**	0.038
4	A, B, C, D, E, F, G, H, **I**	0.033
5	B, C, D, E, F, G, H, I	0.031
6	C, D, E, F, G, H, I	0.030
7	C, E, F, G, H, I	0.027

$$Q(x)=\text{Impurity}(x)+\beta\times\text{Penalty}(k)$$

$$\text{Impurity}(X)=\frac{\#\text{ of Minority Examples}}{n}$$

$$\text{Penalty}(k)=\begin{cases}\sqrt{\frac{k-t}{n}}, & k\geq t\\ 0, & k<t\end{cases}$$



Instituto Nacional de Estatística
Statistics Portugal

SUWAN ALGORITHM

Methodology

- Weighted metric, α

$$Q_{\theta}(P_k^{\alpha})=1-\frac{W_{\theta}(P_k^{\alpha})}{W_{\theta}(P_1)}, \theta\in[1,2]$$

Structural
Compositional
of 68%
of 58%
Qnorm
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
based on D1
based on D2
α=0
α=0.7
α=1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

SUWAN ALGORITHM

Evaluation Measures

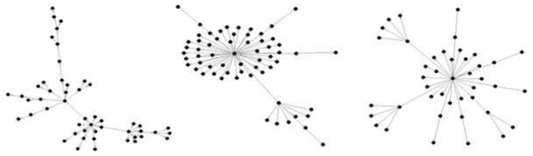
- The implemented methodology works around labeled data hence, the cluster evaluation can be accomplished through the purity of the clusters.
- This way, a new measure of the overall quality based on the cluster's purity is determined to achieve the quality of the clustering.

$$\text{Purity}(C_k) = \max_i (PR_k(t_i))$$

Where $PR_k(t_i)$ is the proportion of class t_i in cluster C_k

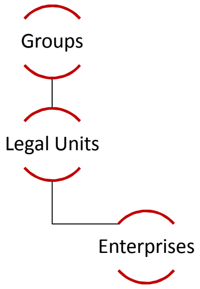
$$\text{Purity}_{\text{total}}(C) = \sum_{k=1}^{j} \frac{|C_k|}{|C|} \times \text{Purity}(C_k)$$

EGR Network



Data

- The database to be explored is provided by INE (Statistics Portugal) and it is denominated by EuroGroups Register (EGR).
- The EGR is a network of register that contains information about multinational enterprise groups, which have statistically relevant financial and non-financial transnational operations in at least one of the European countries (Eurostat, 2010).



Groups

Legal Units

Enterprises

Attribute	Description
LEU_LEID	ID of the Legal Unit
TYPE	List of type of Legal Unit (Brach or not)
LFORM	List of legal forms of Legal Units
COUNTRY_CODE	List of 2-digit ISO country codes
SIZE_CLASS	Size of the enterprise based on persons employed
TURNOVER_CLASS	Turnover class based on the enterprise turnover values
NACE_DIV	2-digit NACE Rev. 2 activity codes for the main activity of enterprises

IMPLEMENTATION

SUWAN VS Subgroup Discovery (SD)

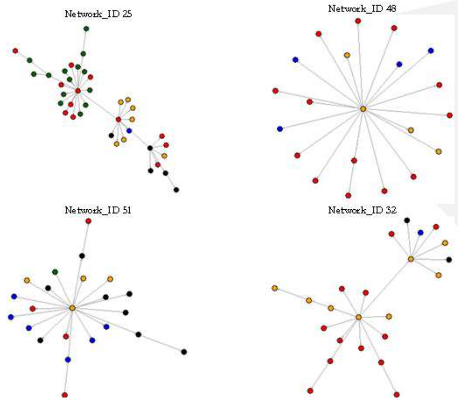
- A Multinational Enterprise Groups represents a network
- Analysis performed on 67 networks with the following characteristics
  - Group Head based in Portugal
  - > 20 connections → # 3 848 Legal Units

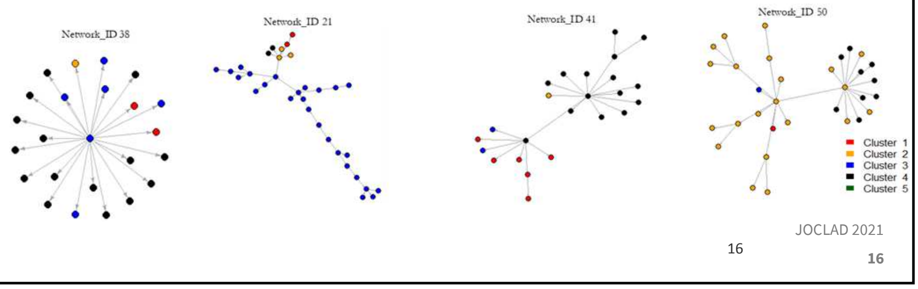
Algorithm	Average #Clusters/Subgroups	Average Overall Quality
SUWAN	3,299	0,532
SD	3,761	0,726

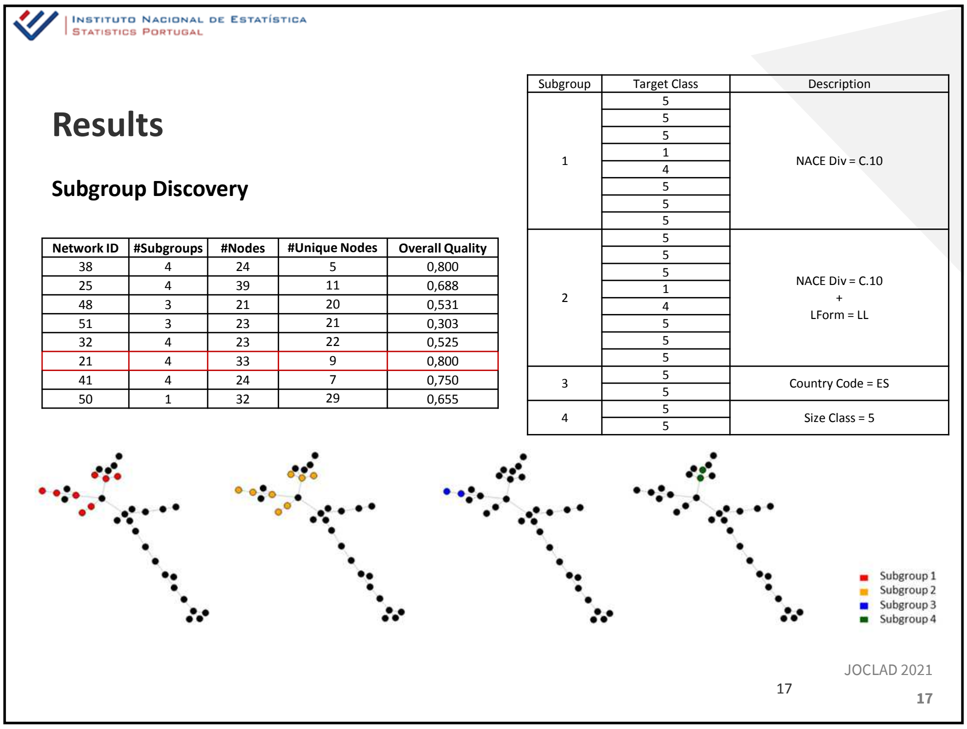
Results

SUWAN

Network ID	#Clusters	#Nodes	Overall Quality
38	4	24	1
25	5	39	0,923
48	3	21	0,905
51	5	23	0,870
32	4	23	0,783
21	4	33	0,758
41	4	24	0,750
50	4	32	0,719







INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

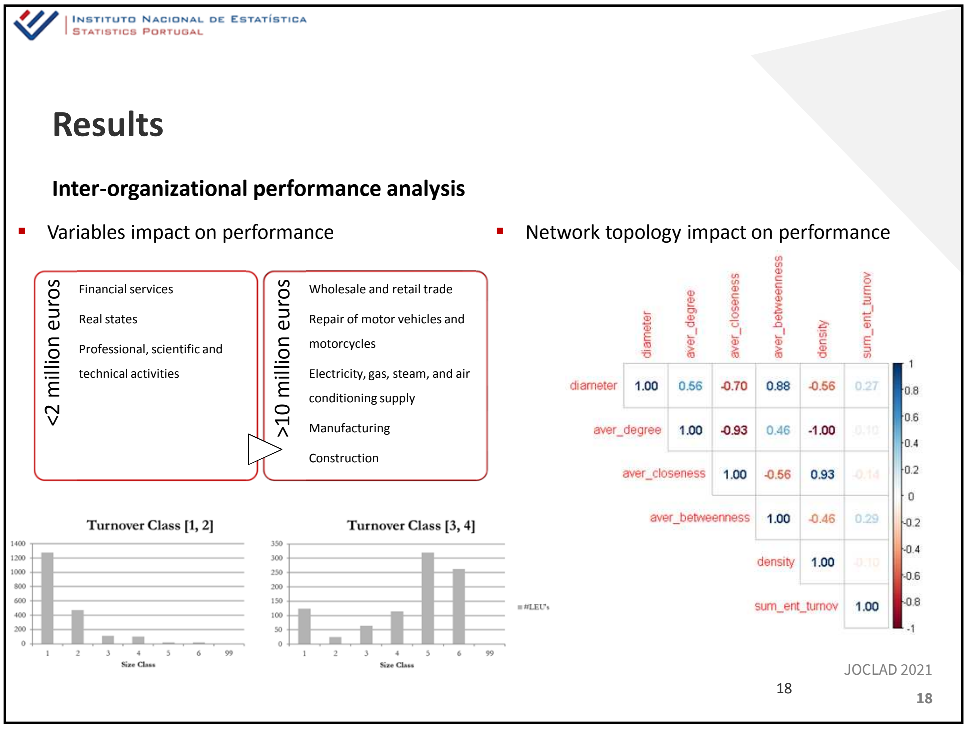
Results

Subgroup Discovery

Network ID	#Subgroups	#Nodes	#Unique Nodes	Overall Quality
38	4	24	5	0,800
25	4	39	11	0,688
48	3	21	20	0,531
51	3	23	21	0,303
32	4	23	22	0,525
21	4	33	9	0,800
41	4	24	7	0,750
50	1	32	29	0,655

Subgroup	Target Class	Description
1	5	NACE Div = C.10
	5	
	5	
	1	
	4	
	5	
	5	
	5	
2	5	NACE Div = C.10 + LForm = LL
	5	
	5	
	1	
	4	
	5	
	5	
	5	
3	5	Country Code = ES
	5	
4	5	Size Class = 5
	5	

- 🟥 Subgroup 1
- 🟧 Subgroup 2
- 🟦 Subgroup 3
- 🟩 Subgroup 4



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Results

Inter-organizational performance analysis

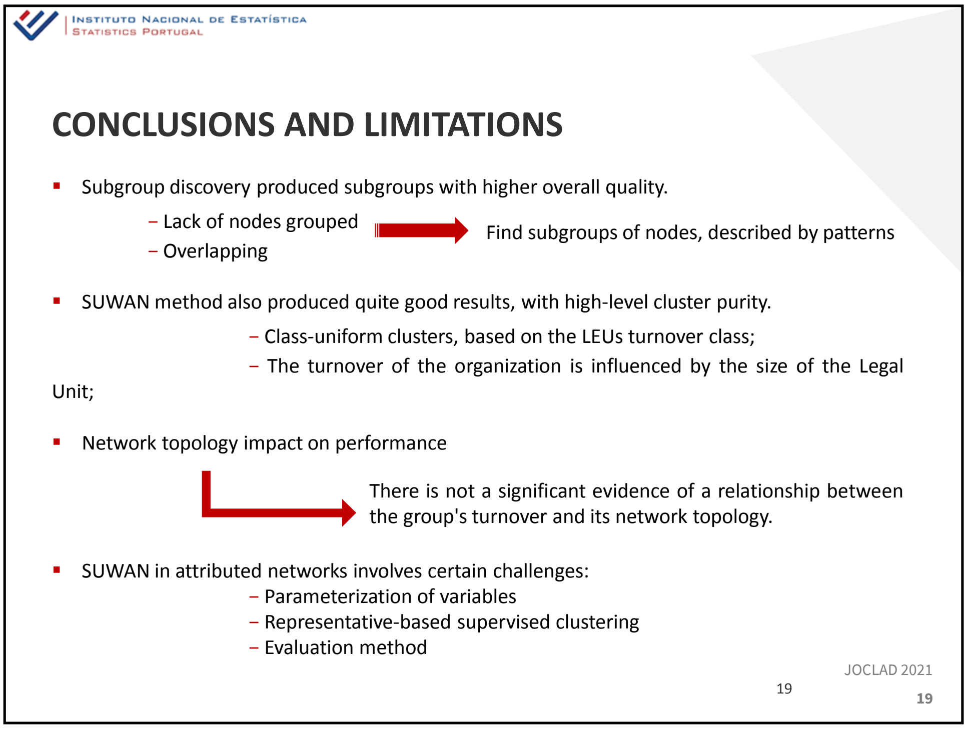
- Variables impact on performance

<2 million euros
Financial services
Real states
Professional, scientific and technical activities

>10 million euros
Wholesale and retail trade
Repair of motor vehicles and motorcycles
Electricity, gas, steam, and air conditioning supply
Manufacturing
Construction

- Network topology impact on performance

	diameter	aver_degree	aver_closeness	aver_betweenness	density	sum_ent_turnov
diameter	1.00	0.56	-0.70	0.88	-0.56	0.27
aver_degree		1.00	-0.93	0.46	-1.00	0.10
aver_closeness			1.00	-0.56	0.93	-0.14
aver_betweenness				1.00	-0.46	0.29
density					1.00	-0.10
sum_ent_turnov						1.00



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CONCLUSIONS AND LIMITATIONS

- Subgroup discovery produced subgroups with higher overall quality.
  - Lack of nodes grouped
  - Overlapping

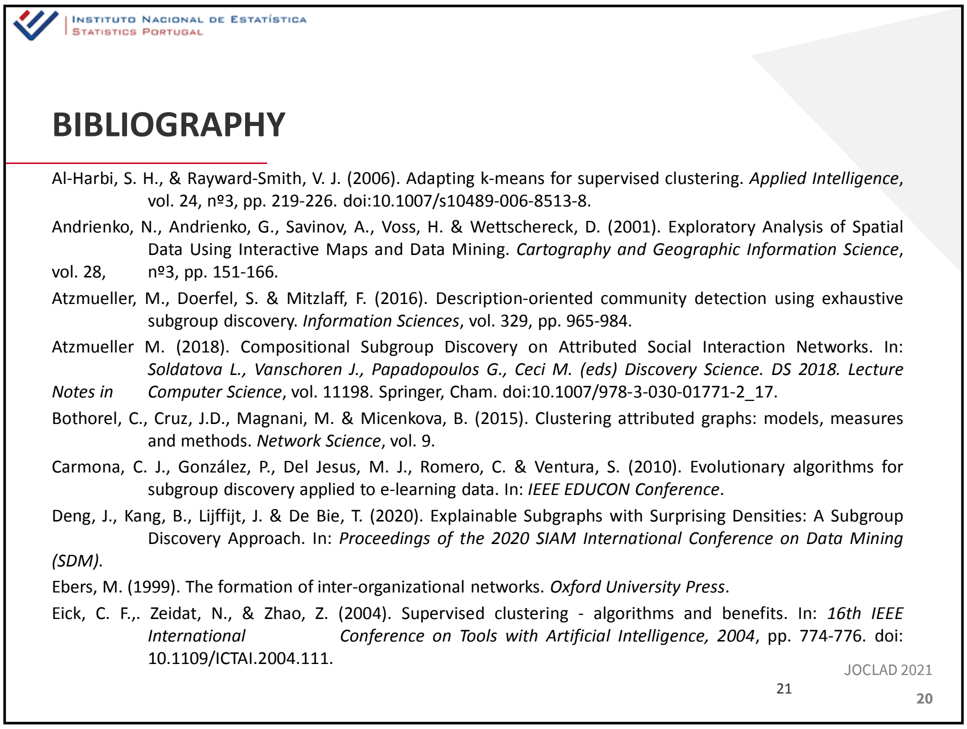
  → Find subgroups of nodes, described by patterns

- SUWAN method also produced quite good results, with high-level cluster purity.
  - Class-uniform clusters, based on the LEUs turnover class;
  - The turnover of the organization is influenced by the size of the Legal Unit;

- Network topology impact on performance

  → There is not a significant evidence of a relationship between the group's turnover and its network topology.

- SUWAN in attributed networks involves certain challenges:
  - Parameterization of variables
  - Representative-based supervised clustering
  - Evaluation method



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

BIBLIOGRAPHY

Al-Harbi, S. H., & Rayward-Smith, V. J. (2006). Adapting k-means for supervised clustering. *Applied Intelligence*, vol. 24, nº3, pp. 219-226. doi:10.1007/s10489-006-8513-8.

Andrienko, N., Andrienko, G., Savinov, A., Voss, H. & Wettschereck, D. (2001). Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining. *Cartography and Geographic Information Science*, vol. 28, nº3, pp. 151-166.

Atzmueller, M., Doerfel, S. & Mitzlaff, F. (2016). Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, vol. 329, pp. 965-984.

Atzmueller M. (2018). Compositional Subgroup Discovery on Attributed Social Interaction Networks. In: *Soldatova L., Vanschoren J., Papadopoulos G., Ceci M. (eds) Discovery Science. DS 2018. Lecture Notes in Computer Science*, vol. 11198. Springer, Cham. doi:10.1007/978-3-030-01771-2_17.

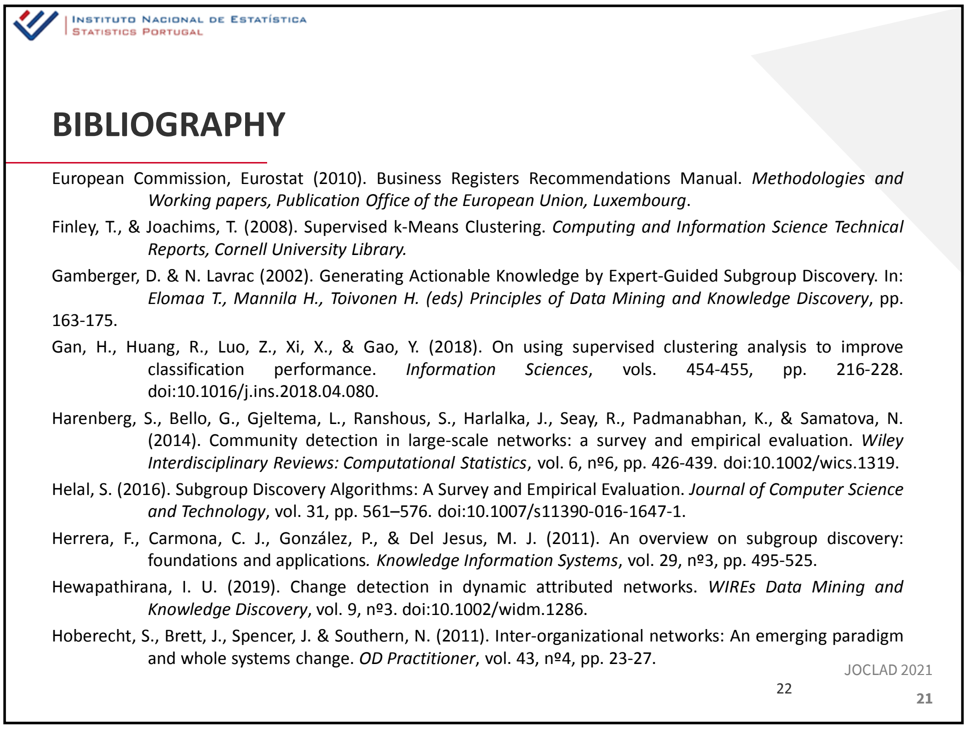
Bothorel, C., Cruz, J.D., Magnani, M. & Micenkova, B. (2015). Clustering attributed graphs: models, measures and methods. *Network Science*, vol. 9.

Carmona, C. J., González, P., Del Jesus, M. J., Romero, C. & Ventura, S. (2010). Evolutionary algorithms for subgroup discovery applied to e-learning data. In: *IEEE EDUCON Conference*.

Deng, J., Kang, B., Lijffijt, J. & De Bie, T. (2020). Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach. In: *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*.

Ebers, M. (1999). The formation of inter-organizational networks. *Oxford University Press*.

Eick, C. F.,. Zeidat, N., & Zhao, Z. (2004). Supervised clustering - algorithms and benefits. In: *16th IEEE International Conference on Tools with Artificial Intelligence, 2004*, pp. 774-776. doi: 10.1109/ICTAI.2004.111.



Instituto Nacional de Estatística
Statistics Portugal

BIBLIOGRAPHY

European Commission, Eurostat (2010). Business Registers Recommendations Manual. *Methodologies and Working papers, Publication Office of the European Union, Luxembourg*.

Finley, T., & Joachims, T. (2008). Supervised k-Means Clustering. *Computing and Information Science Technical Reports, Cornell University Library.*

Gamberger, D. & N. Lavrac (2002). Generating Actionable Knowledge by Expert-Guided Subgroup Discovery. In: *Elomaa T., Mannila H., Toivonen H. (eds) Principles of Data Mining and Knowledge Discovery*, pp. 163-175.

Gan, H., Huang, R., Luo, Z., Xi, X., & Gao, Y. (2018). On using supervised clustering analysis to improve classification performance. *Information Sciences*, vols. 454-455, pp. 216-228. doi:10.1016/j.ins.2018.04.080.

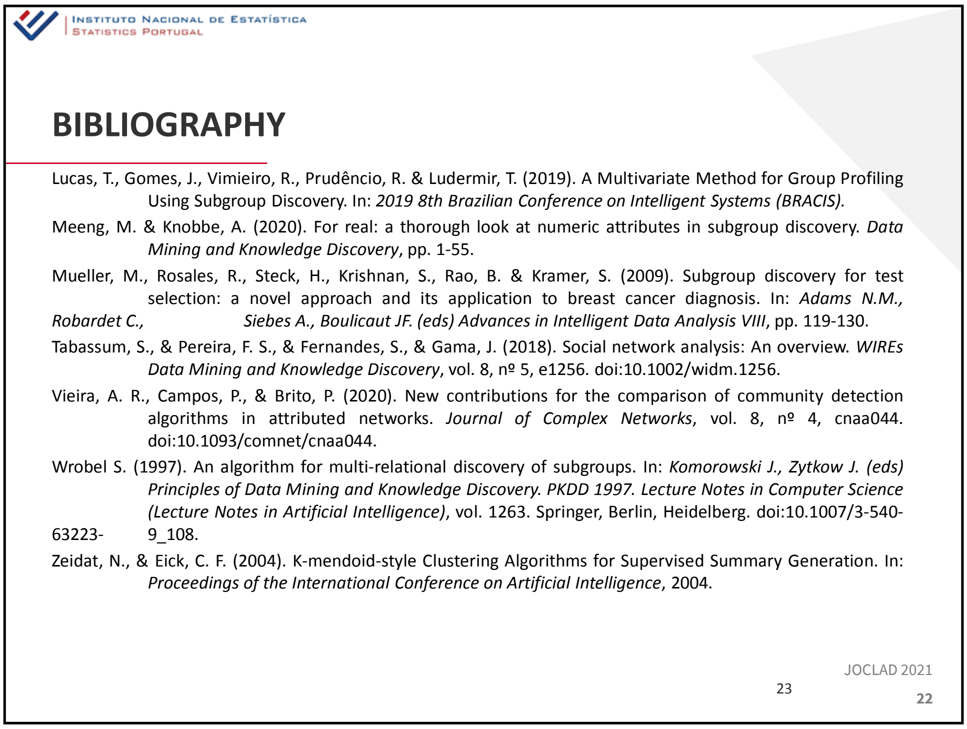
Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., & Samatova, N. (2014). Community detection in large-scale networks: a survey and empirical evaluation. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, nº6, pp. 426-439. doi:10.1002/wics.1319.

Helal, S. (2016). Subgroup Discovery Algorithms: A Survey and Empirical Evaluation. *Journal of Computer Science and Technology*, vol. 31, pp. 561–576. doi:10.1007/s11390-016-1647-1.

Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: foundations and applications. *Knowledge Information Systems*, vol. 29, nº3, pp. 495-525.

Hewapathirana, I. U. (2019). Change detection in dynamic attributed networks. *WIREs Data Mining and Knowledge Discovery*, vol. 9, nº3. doi:10.1002/widm.1286.

Hoberecht, S., Brett, J., Spencer, J. & Southern, N. (2011). Inter-organizational networks: An emerging paradigm and whole systems change. *OD Practitioner*, vol. 43, nº4, pp. 23-27.



Instituto Nacional de Estatística
Statistics Portugal

BIBLIOGRAPHY

Lucas, T., Gomes, J., Vimieiro, R., Prudêncio, R. & Ludermir, T. (2019). A Multivariate Method for Group Profiling Using Subgroup Discovery. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS).*

Meeng, M. & Knobbe, A. (2020). For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, pp. 1-55.

Mueller, M., Rosales, R., Steck, H., Krishnan, S., Rao, B. & Kramer, S. (2009). Subgroup discovery for test selection: a novel approach and its application to breast cancer diagnosis. In: *Adams N.M., Robardet C., Siebes A., Boulicaut JF. (eds) Advances in Intelligent Data Analysis VIII*, pp. 119-130.

Tabassum, S., & Pereira, F. S., & Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. *WIREs Data Mining and Knowledge Discovery*, vol. 8, nº 5, e1256. doi:10.1002/widm.1256.

Vieira, A. R., Campos, P., & Brito, P. (2020). New contributions for the comparison of community detection algorithms in attributed networks. *Journal of Complex Networks*, vol. 8, nº 4, cnaa044. doi:10.1093/comnet/cnaa044.

Wrobel S. (1997). An algorithm for multi-relational discovery of subgroups. In: *Komorowski J., Zytkow J. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 1997. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, vol. 1263. Springer, Berlin, Heidelberg. doi:10.1007/3-540-63223-9_108.

Zeidat, N., & Eick, C. F. (2004). K-mendoid-style Clustering Algorithms for Supervised Summary Generation. In: *Proceedings of the International Conference on Artificial Intelligence*, 2004.