A multiplex-network based approach for clustering ensemble selection

Parisa Rastin & Rushed Kanawati
LIPN, CNRS UMR 7030; USPC
surname.name@lipn.univ-paris13.fr

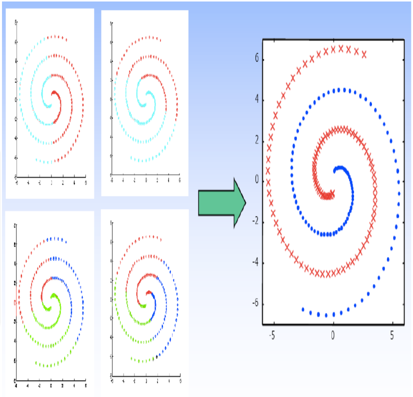
MANEM workshop@ASONAM Paris , 25 August 2015

PLAN

ENSEMBLE CLUSTERING (EC)

Problem definition

- $V = \{v_1, \ldots, v_n\}$
- $\Pi \subset \mathcal{P}(V)$: *base partitions*
- $EC(\Pi) = \pi_* \leftarrow \underset{\pi_i \in \Pi}{\arg\min}\, dist(\pi_*, \pi_i)$



from A. Topchy et. al. Clustering Ensembles: Models of Consensus and Weak Partitions. PAMI, 2005

Applying EC to community detection

- Computing communities cores

  [SG12]
- Dynamic communities

  [LF12]
- Multi-objective (local) community identification

  [Kan15]
- Community detection in multiplex networks

  [FHK14]
- **Yasca** : from local communities to global communities

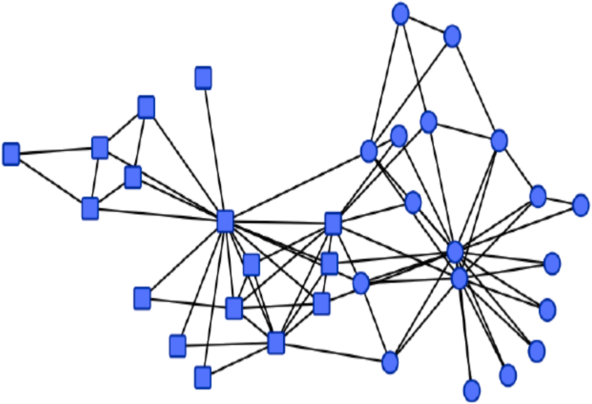
  [Kan14b]
- **Large-scale graph coarsening**

  [OGS10, Ove13, SM13]

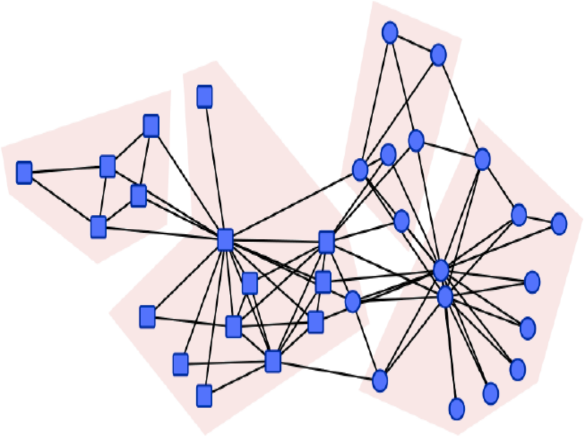
Graph Coarsening

1. Apply N times a fast community detection to the target graph G
   Ex. Applying Label propagation : $\mathcal{O}(m)$
2. Compute the **absolute consensus clustering**.
3. Reduce the graph according to obtained consensus clustering.
4. Apply a high quality community detection algorithm on reduced graph.
5. Expand obtained results to the initial graph.

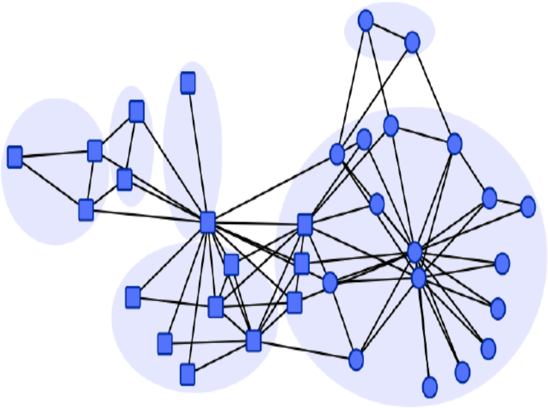
GRAPH COARSENING : ILLUSTRATION I



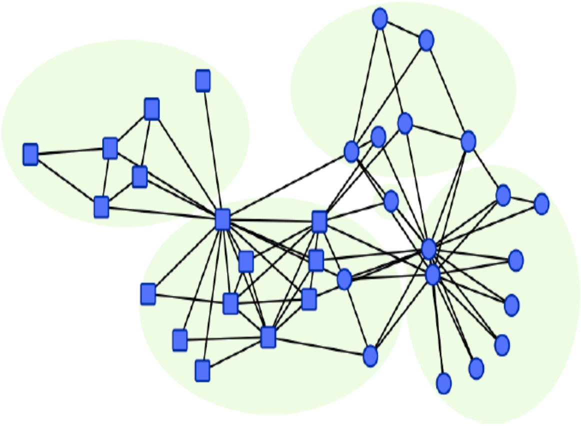
Graph Coarsening : illustration II



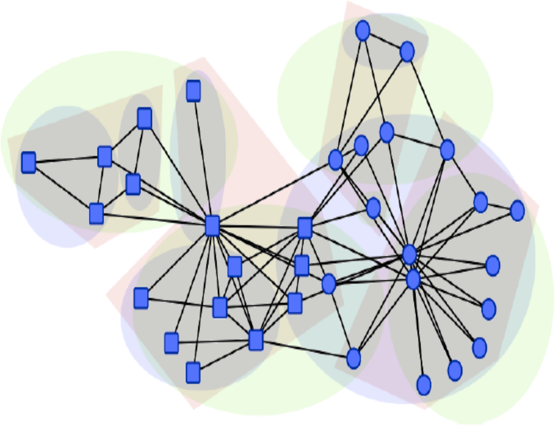
GRAPH COARSENING : ILLUSTRATION III



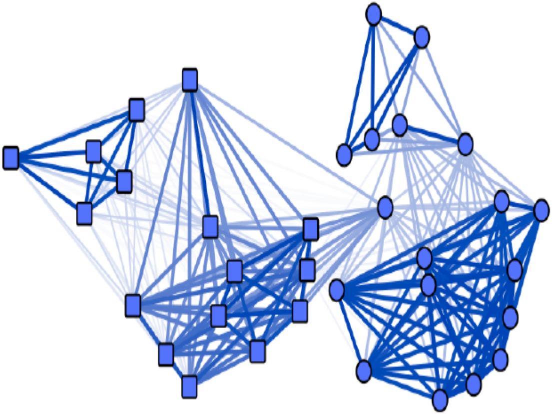
Graph Coarsening : illustration IV



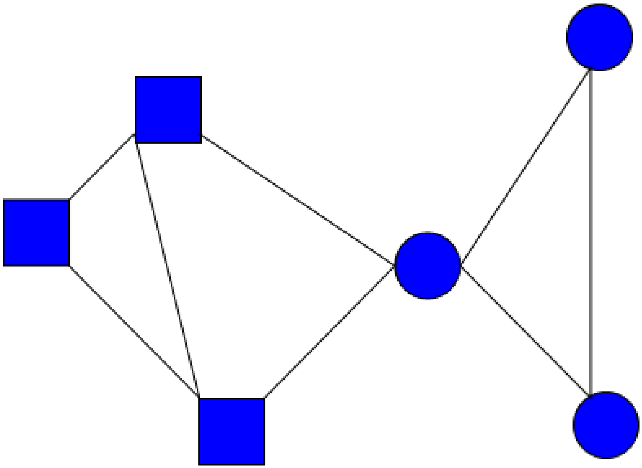
GRAPH COARSENING : ILLUSTRATION V



Graph Coarsening : illustration VI



Graph Coarsening : illustration VII



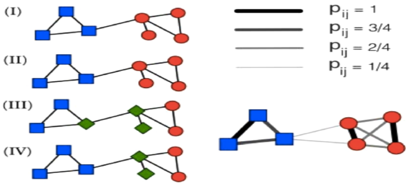
ENSEMBLE CLUSTERING APPROACHES [SG03]

- **CSPA: Cluster-based Similarity Partitioning Algorithm** $\mathcal{O}(n^2kr)$
- HGPA: HyperGraph-Partitioning Algorithm $\mathcal{O}(nkr)$
- MCLA: Meta-Clustering Algorithm $\mathcal{O}(nk^2r^2)$

n # objects, k # of clusters, r# of clustering solutions

CSPA illustration

from [LF12]



ENSEMBLE SELECTION (ES)

Motivation

The quality of a consensus clustering depends on both the **quality** and **diversity** of input base clusterings [FL08, AF09, NCC13, ADIA15].

Problem definition

- Let $\Pi = \{\pi_1, \ldots, \pi_n\}$ be a set of base partitions
- $\mathcal{ES}(\Pi) = \Pi^* \subset \Pi : \mathcal{Q}(EC(\Pi^*)) > \mathcal{Q}(EC(\Pi))$
- $\mathcal{Q}$: Quality of the consensus clustering

Diversity

Clustering Similarity measures

- Purity
- Rand/ARI
- NMI (Normlized mutual information)
- IV (Information variation) [Mei03]
- ...

QUALITY

Cluster internal quality indexes [AR14]

- Silhouette index,
- Calinski-Harabasz index
- Davis-Bouldin index
- Dunn index
- ...

Network-oriented indexes

- Modularity
- Average conductance
- Average local Modularities : L, M, R [Kan15]
- See also [YL12]
- ...

Ensemble selection approaches : limitations

- Existing approaches are defined for attribute/value datasets with metric distances
- Use of one quality/diversity measure.
- Requires the number of clusters to select as input.
- ...

Proposed approach: contributions

- Designed for both networks and attribute/value datasets
- Use of an *ensemble* of quality/diversity measures.
- The number of selected base clustering is automatically computed.

ENSEMBLE SELECTION APPROACH

The idea

- Cluster the set of base clusterings using an ensemble of similarity measures

  Apply a **multiplex community detection** *algorithm to a multiplex network whose nodes are the set of base clusterings and whose layers are defined by a set of* **proximity graphs**, *each defined according a to a given similarity measure*

- From each cluster select the node (i.e clustering) that is ranked first according to an ensemble of quality measures.

  Apply **ensemble ranking** *algorithms*

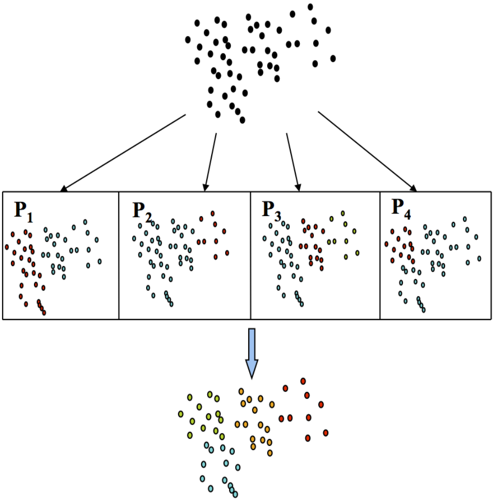
ENSEMBLE SELECTION APPROACH

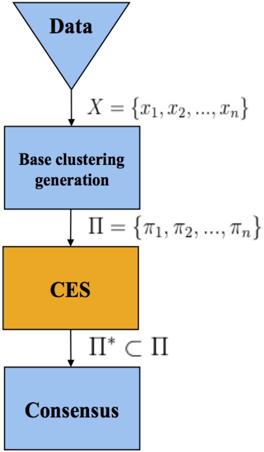
Algorithm 1 Graph-based cluster ensemble selection algorithm

Require: $\Pi = \{\pi_1, \ldots, \pi_r\}$ a set of base clusterings
Require: $\mathcal{S} = \{S_1, \ldots, S_n\}$ A set of partition similarity functions
Require: $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ A set of partition quality functions

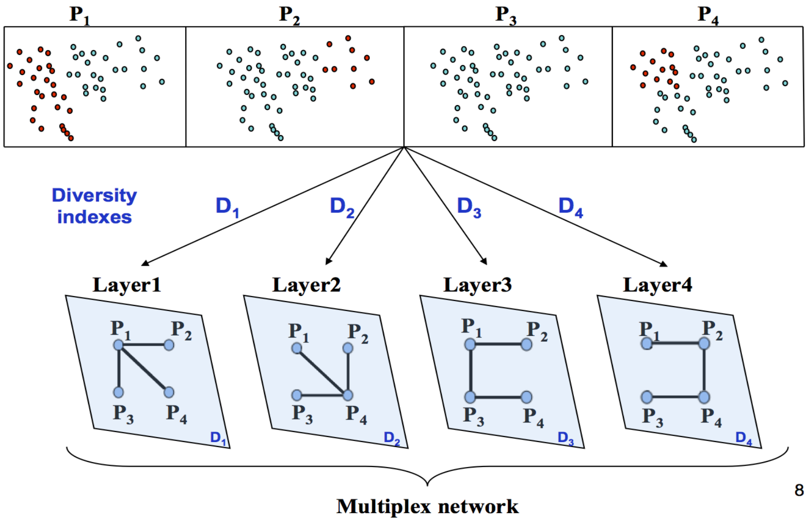
```
Π* ← ∅
MUX ← Multiplex(Π)
for all S_i ∈ 𝒮 do
    MUX.add_layer(proximity_graph(Π, S_i))
end for
𝒞 = {c_1, ..., c_k} ← community_detection(MUX)
for all c ∈ 𝒞 do
    π̂ ← ensemble_Ranking(c, 𝒬)
    Π* ← Π* ∪ {π̂}
end for
return Π*
```

THE PROPOSED APPROACH

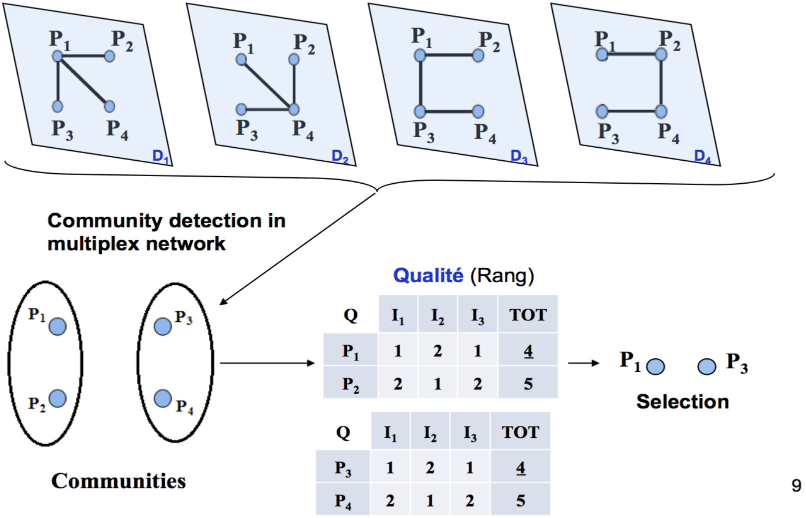




THE PROPOSED APPROACH



The proposed approach



Q	I_1	I_2	I_3	TOT
P_1	1	2	1	4
P_2	2	1	2	5

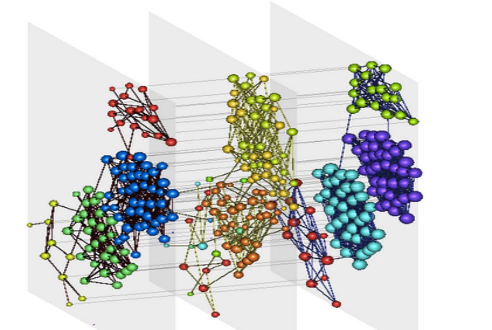
Q	I_1	I_2	I_3	TOT
P_3	1	2	1	4
P_4	2	1	2	5

PROXIMITY GRAPHS

- **ϵ-neighborhood graph** : u, v are linked if $d(u, v) \leq \epsilon$
- **k-nearest neighbor graph** : each node is connected to k nearest nodes.
- **Relative neighborhood graph** :

u, v are linked if $d(u, v) \leq \max_x\{d(v, x), d(u, x)\}, \forall x \neq u, v$

Multiplex network



Multiplex network: A set of nodes related by different types of relations.

COMMUNITY DETECTION IN MULTIPLEX NETWORKS

Approaches [FHK14]

1. **Transformation into a monoplex community detection problem**
   - Layer aggregation approaches [BCG11]
   - Hypergraph transformation based approaches
   - Partition aggregation approaches (Ensemble clustering)
   - Multi-objective approaches [AP14]

2. **Generalization of monoplex oriented algorithms to multiplex networks [MRM$^+$10].**

Applied algorithm : **MuxLicod** a seed-centric algorithm [HK15]

Seed-centric algorithms [Kan14a]

Algorithm 2 General seed-centric community detection algorithm

Require: $G =< V, E >$ a connected graph,

$\mathcal{C} \leftarrow \emptyset$

$S \leftarrow$ **compute_seeds(G)**

for $s \in S$ **do**

  $C_s \leftarrow$ **compute_local_com(s,G)**

  $\mathcal{C} \leftarrow \mathcal{C} + C_s$

end for

return compute_community($\mathcal{C}$)

ENSEMBLE RANKING

Problem

- Let L be a set of elements to rank by n rankers
- Let σ_i be the rank provided by ranker i
- **Goal: Compute a consensus rank of** L.

Déjà Vu: Social choice algorithms, but . . .

- Small number of voters and big number of candidates
- Algorithmic efficiency is required

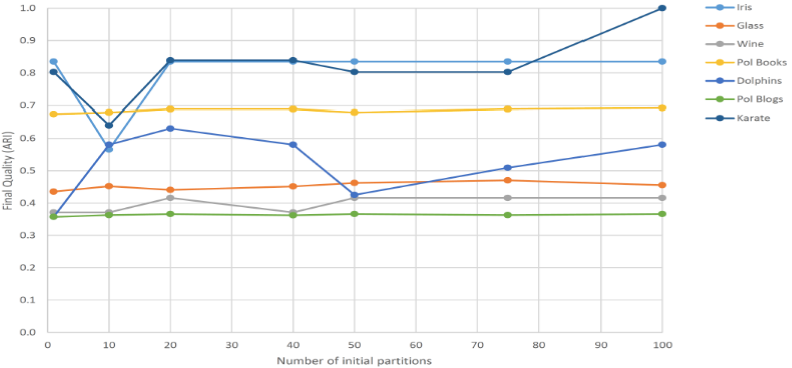
Algorithms

- Borda
- Kemeny approaches (computing Condorcet winner if it exists)

EXPERIMENT ON SMALL NETWORKS WITH KNOWN GROUND-TRUTH PARTITIONS

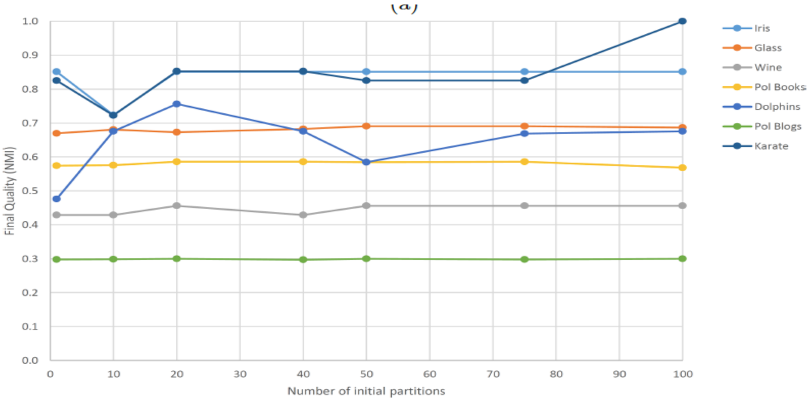
- Small benchmark networks : Karate club, Poltical books, Political blogs, Dolphins
- UCI datasets : Iris, Wine, Glass (transformed into into networks applying RNG)
- Variation of number of base clusterings [1, 100]
- Evaluation of output in function of NMI, ARI.

EXPERIMENT ON SMALL NETWORKS WITH KNOWN GROUND-TRUTH PARTITIONS



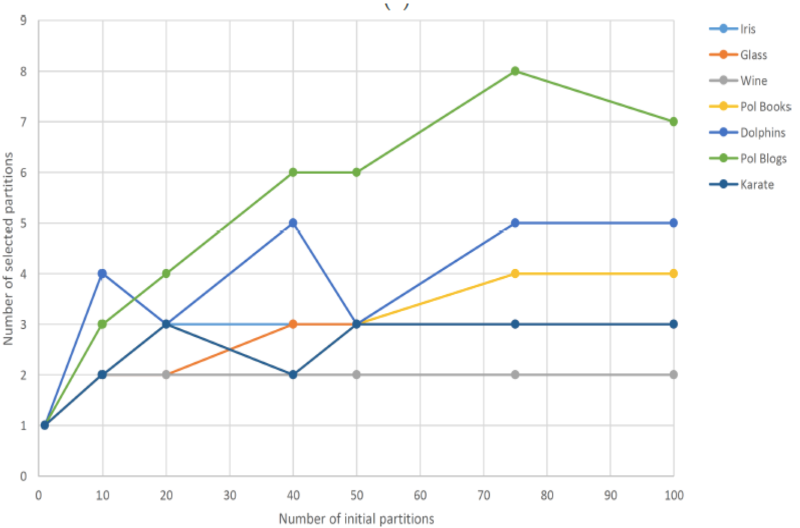
Quality of the selection (ARI) / # base partitions

EXPERIMENT ON SMALL NETWORKS WITH KNOWN GROUND-TRUTH PARTITIONS

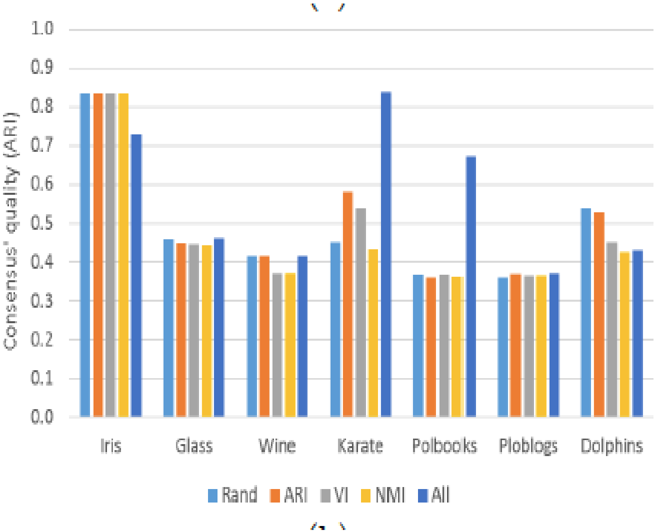


Quality of the selection (ARI) / # base partitions

Experiment on small networks with known ground-truth partitions



EXPERIMENT ON SMALL NETWORKS WITH KNOWN GROUND-TRUTH PARTITIONS



EXPERIMENT II : DBLP CO-AUTHORSHIP NETWORK

- Co-authorship network 1970-1977 (GCC) : $|V| = 643$, $|m| = 886$
- Generation of 10, 100 base clusterings
- Proximity graphs : RNG
- $\mathcal{S} = \{$ NMI, ARI, VI $\}$ $\mathcal{Q} = \{$ modularity, Local modularities L, M, R $\}$

Table: Evaluation of the proposed graph-based ensemble selection

# base clusterings	10
Nodes Compression without selection	18,3%
Nodes Compression with selection	20,9%
Edge compression without selection	17,2%
Edge compression with selection	17,6%
Modularity without selection	0.3734
Modularity with selection	**0.43756**

Experiment II : DBLP co-authorship network

Table: Evaluation of the proposed graph-based ensemble selection

# base clusterings	100
Nodes Compression without selection	35,1%
Nodes Compression with selection	40,3%
Edge compression without selection	36,2%
Edge compression with selection	38,3%
Modularity without selection	0.4031
Modularity with selection	**0.4665**

CONCLUSION & FUTURE WORK

Conclusion

- A new approach for ensemble selection
- The approach can be applied to both networks and attribute/value datasets clustering
- Ensemble selection enhances both the compression ratio and the quality of reduced graphs.

Underwork

- Evaluation on large-scale graphs
- Task oriented evaluation : Recommender systems
  Tag recommendations & Movie rating tasks
- Study of effects of the choice of : proximity graph, multiplex community detection algorithm and choice of the consensus function.

That's all folks !

Questions ?

BIBLIOGRAPHY I

Ebrahim Akbari, Halina Mohamed Dahlan, Roliana Ibrahim, and Hosein Alizadeh, *Hierarchical cluster ensemble selection*, Engineering Applications of Artificial Intelligence **39** (2015), 146–156.

Javad Azimi and Xiaoli Fern, *Adaptive cluster ensemble selection*, IJCAI (Craig Boutilier, ed.), 2009, pp. 992–997.

Alessia Amelio and Clara Pizzuti, *Community detection in multidimensional networks*, IEEE 26th International Conference on Tools with Artificial Intelligence, 2014, pp. 352–359.

Charu C. Aggarwal and Chandan K. Reddy (eds.), *Data clustering: Algorithms and applications*, CRC Press, 2014.

Michele Berlingerio, Michele Coscia, and Fosca Giannotti, *Finding and characterizing communities in multidimensional networks*, ASONAM, IEEE Computer Society, 2011, pp. 490–494.

Bibliography II

Issam Falih, Manel Hmimida, and Rushed Kanawati, *Community detection in multiplex network: a comparative study*, Proceedings of Multiplex networks, Satelitte workshop at European conference on complex systems (Lucca, Italy), September 2014.

Xiaoli Z. Fern and Wei Lin, *Cluster ensemble selection*, Statistical Analysis and Data Mining **1** (2008), no. 3, 128–141.

Manel Hmimida and Rushed Kanawati, *Community detection in multiplex networks: A seed-centric approach*, Networks and Heterogeneous Media **10** (2015), no. 1, 71–85, Special Issue on New trends, models and applications in Complex and Multiplex Networks.

Rushed Kanawati, *Seed-centric approaches for community detection in complex networks*, 6th international conference on Social Computing and Social Media (Crete, Greece) (Gabriele Meiselwitz, ed.), vol. LNCS 8531, Springer, June 2014, pp. 197–208.

______, *Yasca: An ensemble-based approach for community detection in complex networks*, COCOON (Zhipeng Cai, Alex Zelikovsky, and Anu G. Bourgeois, eds.), Lecture Notes in Computer Science, vol. 8591, Springer, 2014, pp. 657–666.

Bibliography III

______, *Empirical evaluation of applying ensemble methods to ego-centered community identification in complex networks*, Neurocomputing **150, B** (2015), 417–427.

Andrea Lancichinetti and Santo Fortunato, *Consensus clustering in complex networks*, Sci. Rep. **2** (2012).

Marina Meila, *Comparing clusterings by the variation of information*, COLT (Bernhard Schölkopf and Manfred K. Warmuth, eds.), Lecture Notes in Computer Science, vol. 2777, Springer, 2003, pp. 173–187.

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela, *Community structure in time-dependent, multiscale, and multiplex networks*, Science **328** (2010), no. 5980, 876–878.

Murilo Coelho Naldi, André C. P. L. F. Carvalho, and Ricardo J. G. B. Campello, *Cluster ensemble selection based on relative validity indexes*, Data Min. Knowl. Discov. **27** (2013), no. 2, 259–289.

BIBLIOGRAPHY IV

- Michael Ovelgönne and Andreas Geyer-Schulz, *Cluster cores and modularity maximization*, ICDM Workshops, 2010, pp. 1204–1213.
- Michael Ovelgönne, *Distributed community detection in web-scale networks*, ASONAM (Jon G. Rokne and Christos Faloutsos, eds.), ACM, 2013, pp. 66–73.
- A. Strehl and J. Ghosh, *Cluster ensembles: a knowledge reuse framework for combining multiple partitions*, The Journal of Machine Learning Research **3** (2003), 583–617.
- Massoud Seifi and Jean-Loup Guillaume, *Community cores in evolving networks*, WWW (Companion Volume) (Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, eds.), ACM, 2012, pp. 1173–1180.
- Christian Staudt and Henning Meyerhenke, *Engineering high-performance community detection heuristics for massive graphs*, ICPP, IEEE, 2013, pp. 180–189.

BIBLIOGRAPHY V

Jaewon Yang and Jure Leskovec, *Defining and evaluating network communities based on ground-truth*, ICDM (Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, eds.), IEEE Computer Society, 2012, pp. 745–754.