Are You My Neighbor? Bringing Order to Neighbor Computing Problems.

David C. Anastasiu[1,2] , Huzefa Rangwala[3], and Andrea Tagarelli[4]

[1]Computer Engineering, San Jose State University, CA

[1]Computer Science & Engineering, Santa Clara University, CA

[2]Computer Science & Engineering, George Mason University, VA

[3]DIMES, University of Calabria, Italy

25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 2019

Part VI:
Neighbors in Learning and Mining Problems in Graph Data

Andrea Tagarelli, University of Calabria, Italy [andrea.tagarelli@unical.it]

Tutorial Outline

- **Part I: Problems and Data Types**
  - Dense, sparse, and asymmetric data
  - Bounded nearest neighbor search
  - Nearest neighbor graph construction
  - Classical approaches and limitations
- **Part II: Neighbors in Genomics, Proteomics, and Bioinformatics**
  - Mass spectrometry search
  - Microbiome analysis
- **Part III: Approximate Search**
  - Locality sensitive hashing variants
  - Permutation and graph-based search
  - Maximum inner product search
- **Part IV: Neighbors in Advertising and Recommender Systems**
  - Collaborative filtering at scale
  - Learning models based on the neighborhood structure
- **Part V: Filtering-Based Search**
  - Massive search space pruning by partial indexing
  - Effective proximity bounds and when they are most useful
- **Part VI: Neighbors in Learning and Mining Problems in Graph Data**
  - Neighborhood as cluster in a complex network system
  - Neighborhood as influence trigger set

Neighborhood in graphs

- Node neighborhood computation is key-enabling in a variety of graph mining problems
  - Centrality
  - Clustering
  - Community search, detection, evolution
  - Link prediction
  - Information diffusion
  - Influence propagation
  - Representation learning

Neighborhood in graphs

Covered in this part of the tutorial

- Neighborhood as cluster in a complex network system
  - Consensus multilayer community detection from an ensemble of community structures
  - Node-centric (or local) multilayer community detection
- Neighborhood as influence trigger set
  - Community-based (targeted) influence maximization
  - Topology-driven diversity-based (targeted) influence maximization

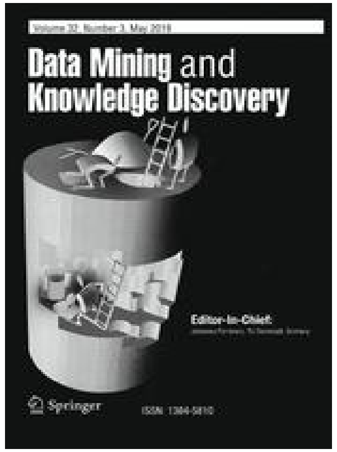
NEIGHBORHOOD AS CLUSTER IN A COMPLEX NETWORK SYSTEM

Consensus multilayer community detection from an ensemble of community structures

Main references for this part

The 2017 European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases

ECML PKDD 2017



A. Tagarelli, A. Amelio, F. Gullo:
Ensemble-based community detection in multilayer networks.
Data Min. Knowl. Discov. 31(5): 1506-1543 (2017)

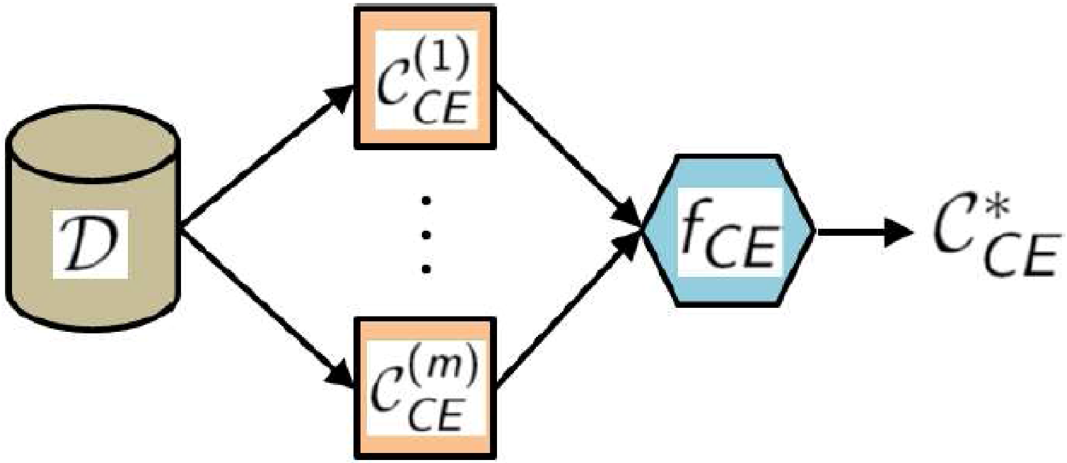
PAKDD 2018

The 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining

June 3rd - 6th, 2018, Melbourne, Australia

D. Mandaglio, A. Amelio, A. Tagarelli:
Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning. In *Proc. PAKDD Conf*. (2018)

Preamble: Ensemble Clustering



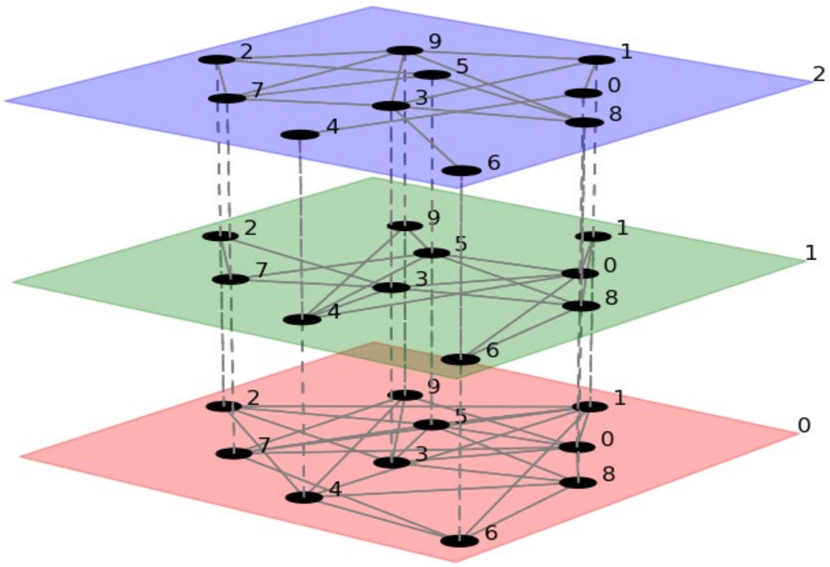
input an *ensemble*, i.e., a set $\mathcal{E}_{CE} = \{\mathcal{C}_{CE}^{(1)}, \ldots, \mathcal{C}_{CE}^{(m)}\}$ of clustering solutions defined over the same set $\mathcal{D}$ of data objects

output a *consensus clustering* $\mathcal{C}_{CE}^{*}$ that aggregates the information from $\mathcal{E}_{CE}$ by optimizing a *consensus function* $f_{CE}(\mathcal{E}_{CE})$

Preamble: Multilayer Network model

Different relations can occur among the same individuals, in different contexts

- Multiple accounts across different online social networks
- Online/offline relationships among the same group of individuals (e.g., followship, like/comment interactions, working together, having lunch) to model complex behaviors



Multilayer network model

- All of the layers share the same population (set of entities)
- Each layer models a different entity relation
- Each entity participates in at least one layer
- *Multiplex* constraint: Inter-layer couplings connect nodes corresponding to the same entity

Motivations

- **Core problem in CD**: to identify the “best” method and its configuration for a target application domain
- Many real-world network systems are *complex*
  - communities can have very different structure and meanings depending on the node relation and/or the dimension/view

- **Key idea**:
  - Model multiple community structures for the same graph to infer robust, high-quality **consensus** results
    - *Not really novel but still largely unexplored*
  - Exploit *multilayer network* model to retain richer/more diverse info than both separate and flattened representations

Early work on consensus community detection

Given a weighted graph G, a selected community detection algorithm A, a desired number of clusterings n_p, and a real-valued threshold θ

1. Apply A on G n_p times to obtain n_p clusterings
2. Build the co-association matrix M (without any constraint on node linkage) and threshold it using θ
3. Apply A on M n_p times
4. If the obtained clusterings are all equal then STOP, else go back to Step 2

Lancichinetti, A., Fortunato, S.: *Consensus clustering in complex networks*. Sci. Rep. 2, 336 (2012)

Consensus community detection as optimization problem

- First well-principled formulation of the **ensemble-based community detection** problem, under multilayer network model
- Aggregation accounts **for intra-community and inter-community connectivity**, besides node membership
- Consensus function is optimized via **multilayer modularity** analysis
- Consensus solution is discovered from a space of candidates delimited by two community structures representative of the ensemble

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Highlights (1/2)

- Two baseline methods using *co-association-based consensus clustering*
  - community structure as **topological upper-bound** and **topological lower-bound**, resp., of the input multilayer network

- New problem of modularity-optimization-driven ensemble-based multilayer community detection (M-EMCD)
  - Consensus solution with maximum modularity
  - Search space of community structures that are
    - valid w.r.t. the input ensemble,
    - and topologically bounded by the baseline solutions

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Highlights (2/2)

- Hill-climbing method for the M-EMCD problem
  - Linear in the number of multilayer edges

- Experimental evaluation based on
  - 7 real-world multilayer networks + *mLFR* benchmark
  - multilayer *modularity*, multilayer *silhouette*, *redundancy*, *NMI*
  - 7 state-of-the-art ML-CD methods

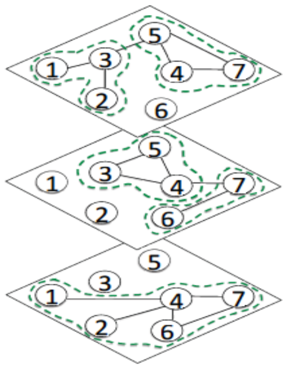
- M-EMCD consensus solution has far better multilayer modularity and quality of community memberships w.r.t. the ensemble-based baseline methods and competing methods

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Ensemble-based baselines

Direct cluster-induced EMCD (C-EMCD)

- Requires construction of the **co-association matrix**
- Subject to within-community **node linkage constraint**
- Pruned according to a given *min co-association* threshold θ



$\theta \geq 2/3$

	1	2	3	4	5	6	7
1	0	2/3	1/3	1/3	0	1/3	1/3
2	2/3	0	1/3	1/3	0	1/3	1/3
3	1/3	1/3	0	1/3	1/3	0	0
4	1/3	1/3	1/3	0	2/3	1/3	2/3
5	0	0	1/3	2/3	0	0	1/3
6	1/3	1/3	0	1/3	0	0	2/3
7	1/3	1/3	0	2/3	1/3	2/3	0

$C_1 = \{1,2\}$
$C_2 = \{3\}$
$C_3 = \{4,5,6,7\}$

	1	2	3	4	5	6	7
1	0	0	1/3	1/3	0	0	0
2	0	0	1/3	1/3	0	0	0
3	1/3	1/3	0	1/3	1/3	0	0
4	1/3	1/3	1/3	0	2/3	1/3	2/3
5	0	0	1/3	2/3	0	0	1/3
6	0	0	0	1/3	0	0	2/3
7	0	0	0	2/3	1/3	2/3	0

$C_1 = \{1\}$ $C_2 = \{2\}$
$C_3 = \{3\}$ $C_4 = \{4,5,6,7\}$

- Consensus community structure via projections on the θ-thresholded matrix
  - Each community corresponds to a multilayer, connected subgraph induced from a cluster

➔ Topological *upper-bound* consensus

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Ensemble-based baselines

Constrained cluster-induced EMCD (CC-EMCD)

- C-EMCD discards the contribution of each specific layer to the node co-associations

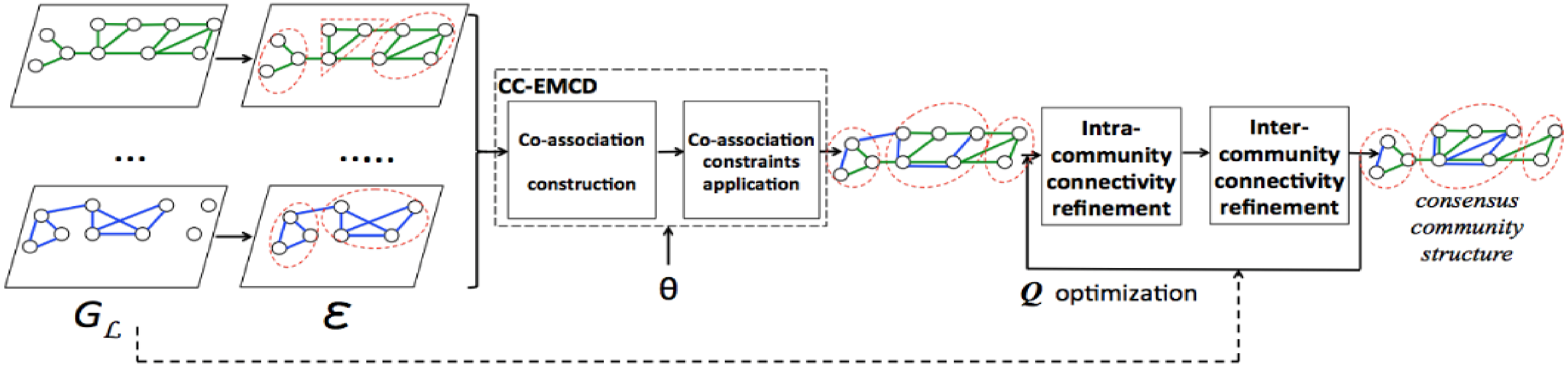
- Two topological refinements:
  - a consensus community is comprised only of edges from layers contributing to the connection of nodes
  - any two consensus communities are linked through edges that correspond to layers in which any two nodes do not appear in the co-association matrix

➔ Topological *lower-bound* consensus

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Modularity-driven EMCD framework

- Consensus solutions by C-EMCD and CC-EMCD might be
  - *redundant*, in terms of multilayer edges connecting different communities (C-EMCD)
  - *poorly descriptive*, in terms of multilayer edges that characterize their internal connectivity of communities (CC-EMCD)
- **Idea:** Find modularity-optimal consensus over the search space delimited by CC-EMCD and C-EMCD solutions


...
.....
$G_{\mathcal{L}}$
$\mathcal{E}$
CC-EMCD
Co-association construction
Co-association constraints application
θ
Intra-community connectivity refinement
Inter-community connectivity refinement
Q optimization
consensus community structure

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Multilayer modularity (1/2)

$$Q(\mathcal{C}) = \sum_{C\in\mathcal{C}} Q(C) = \frac{1}{\underbrace{d(V_{\mathcal{L}})}_{\text{total multilayer degree}}} \sum_{C\in\mathcal{C}} \sum_{L\in\mathcal{L}} \left(\underbrace{d_L^{int}(C)}_{\text{internal degree}} - \underbrace{\gamma_L}_{\text{resolution}} \frac{(d_L(C))^2}{d(V_{\mathcal{L}})} + \overset{\boxed{0/1}}{\beta} \sum_{\underbrace{L'\in\mathcal{P}(L)}_{\text{set of valid layer-pairings}}} \underbrace{d^{ext}_{L,L'}(C)}_{\text{inter-layer coupling}} \right)$$

$$\mathcal{P}(L) = \begin{cases} \{L' \in \mathcal{L} \mid L \prec_{\mathcal{L}} L'\}, & \textit{if } \prec_{\mathcal{L}} \textit{ is defined} \\ \mathcal{L} \setminus \{L\}, & \textit{otherwise} \end{cases}$$

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Multilayer modularity (2/2)

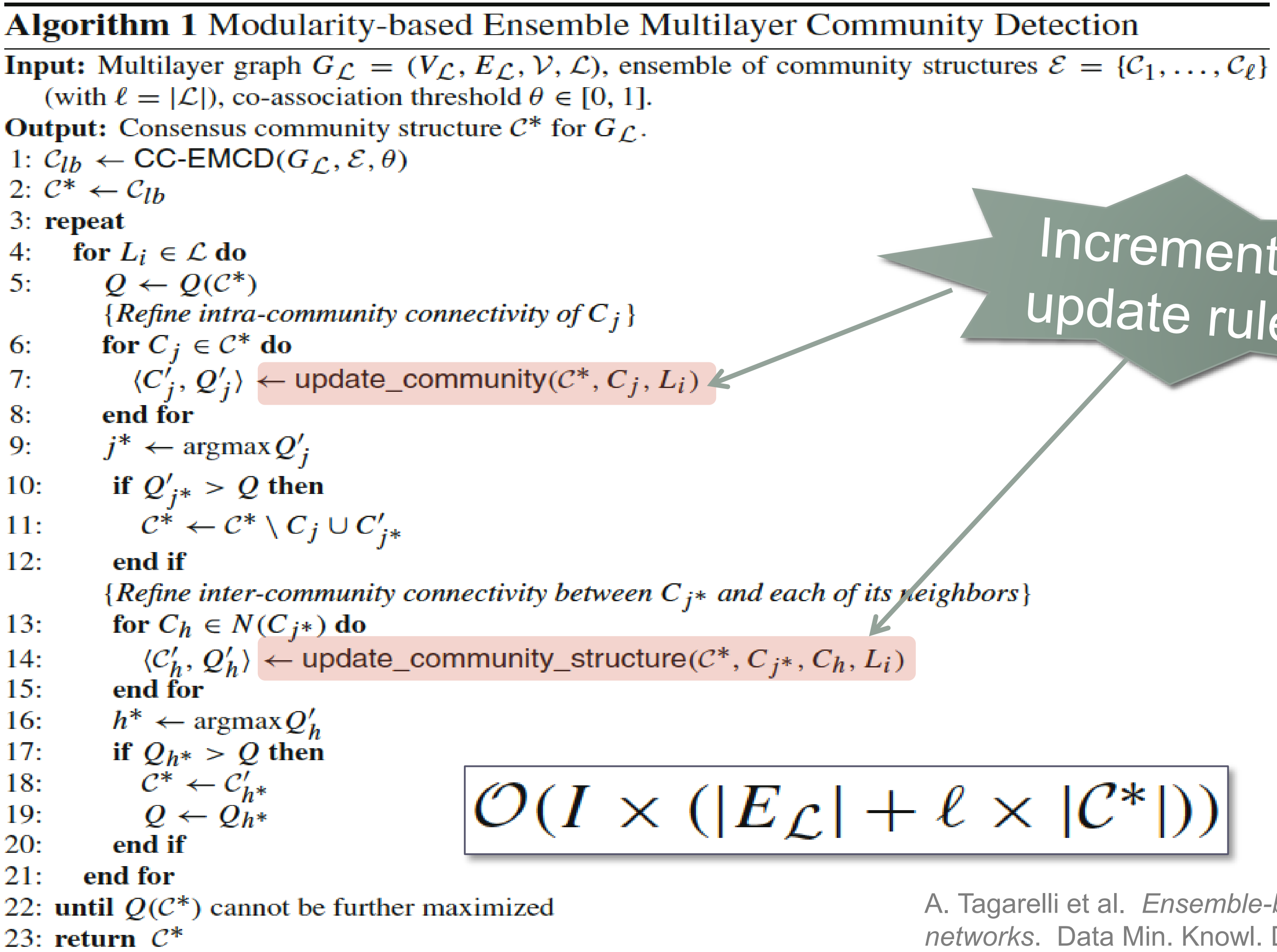
- **Goals**
  - Overcome issues in multislice modularity
  - Provide principled definitions for the resolution and inter-layer coupling factors
  - Manage modularity in time-evolving networks

- Main **contributions**
  - Exploit structure at graph and community level
  - Redundancy-based resolution factor
    - specifically for any given pair of layer and community
  - Projective-based inter-layer coupling factor
    - accounts for properties of a community projection over any two comparable layers
    - Partial order relation over the layers

A. Amelio, G. Mangioni, A. Tagarelli
Modularity in Multilayer Networks using Redundancy-based Resolution and Projection-based Inter-Layer Coupling.
IEEE Trans. Netw. Sci. Eng. (2019)

A. Amelio, A. Tagarelli
Revisiting Resolution and Inter-Layer Coupling Factors in Modularity for Multilayer Networks.
In *Proc. ASONAM* 2017

The **M-EMCD** algorithm



Algorithm 1 Modularity-based Ensemble Multilayer Community Detection

Input: Multilayer graph $G_{\mathcal{L}} = (V_{\mathcal{L}}, E_{\mathcal{L}}, \mathcal{V}, \mathcal{L})$, ensemble of community structures $\mathcal{E} = \{\mathcal{C}_1, \ldots, \mathcal{C}_\ell\}$ (with $\ell = |\mathcal{L}|$), co-association threshold $\theta \in [0, 1]$.

Output: Consensus community structure $\mathcal{C}^*$ for $G_{\mathcal{L}}$.

```
$\mathcal{C}_{lb} \leftarrow$ CC-EMCD$(G_{\mathcal{L}}, \mathcal{E}, \theta)$
$\mathcal{C}^* \leftarrow \mathcal{C}_{lb}$
repeat
  for $L_i \in \mathcal{L}$ do
    $Q \leftarrow Q(\mathcal{C}^*)$
    {Refine intra-community connectivity of $C_j$}
    for $C_j \in \mathcal{C}^*$ do
      $\langle C'_j, Q'_j \rangle \leftarrow$ update_community$(\mathcal{C}^*, C_j, L_i)$
    end for
    $j^* \leftarrow \operatorname{argmax} Q'_j$
    if $Q'_{j*} > Q$ then
      $\mathcal{C}^* \leftarrow \mathcal{C}^* \setminus C_j \cup C'_{j*}$
    end if
    {Refine inter-community connectivity between $C_{j*}$ and each of its neighbors}
    for $C_h \in N(C_{j*})$ do
      $\langle C'_h, Q'_h \rangle \leftarrow$ update_community_structure$(\mathcal{C}^*, C_{j*}, C_h, L_i)$
    end for
    $h^* \leftarrow \operatorname{argmax} Q'_h$
    if $Q_{h*} > Q$ then
      $\mathcal{C}^* \leftarrow \mathcal{C}'_{h*}$
      $Q \leftarrow Q_{h*}$
    end if
  end for
until $Q(\mathcal{C}^*)$ cannot be further maximized
return $\mathcal{C}^*$
```

Incremental update rules

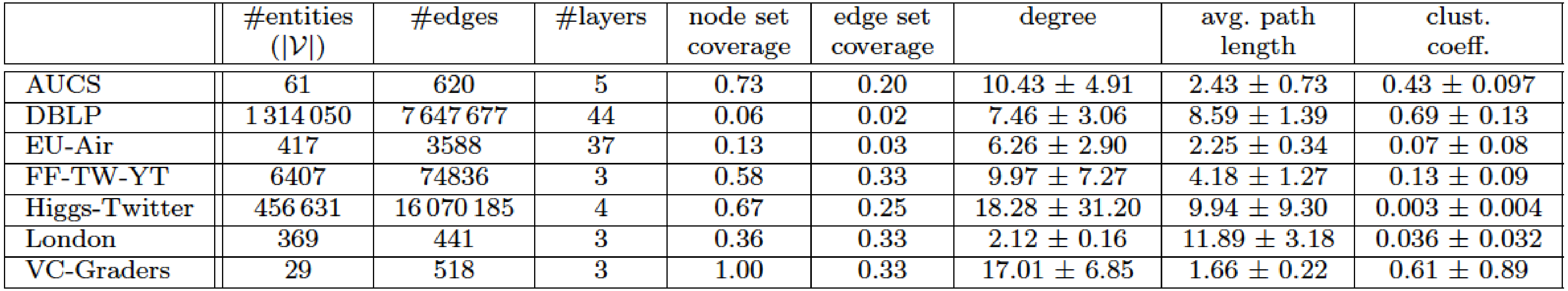
$$\boxed{\mathcal{O}(I \times (|E_{\mathcal{L}}| + \ell \times |\mathcal{C}^*|))}$$

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Experimental evaluation

Datasets

- **Seven real-world** multilayer network datasets



	#entities ($\|\mathcal{V}\|$)	#edges	#layers	node set coverage	edge set coverage	degree	avg. path length	clust. coeff.
AUCS	61	620	5	0.73	0.20	10.43 ± 4.91	2.43 ± 0.73	0.43 ± 0.097
DBLP	1 314 050	7 647 677	44	0.06	0.02	7.46 ± 3.06	8.59 ± 1.39	0.69 ± 0.13
EU-Air	417	3588	37	0.13	0.03	6.26 ± 2.90	2.25 ± 0.34	0.07 ± 0.08
FF-TW-YT	6407	74836	3	0.58	0.33	9.97 ± 7.27	4.18 ± 1.27	0.13 ± 0.09
Higgs-Twitter	456 631	16 070 185	4	0.67	0.25	18.28 ± 31.20	9.94 ± 9.30	0.003 ± 0.004
London	369	441	3	0.36	0.33	2.12 ± 0.16	11.89 ± 3.18	0.036 ± 0.032
VC-Graders	29	518	3	1.00	0.33	17.01 ± 6.85	1.66 ± 0.22	0.61 ± 0.89

- Brodka's **mLFR** to create multilayer network with **1M nodes**
  - Used for efficiency evaluation
  - Parameter setting:
    - 10 layers
    - average degree 30, maximum degree 100
    - mixing at 20%
    - layer mixing 2

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Experimental evaluation

Competing methods

- **Flattening** approach
  - *Nerstrand* [1]
- **Aggregation** approach
  - *Principal Modularity Maximization* (PMM) [2]
  - *frequent pAttern mining-BAsed Community discoverer in mUltidimensional networkS* (ABACUS) [3]
- **Direct** approach
  - *Generalized Louvain* (GL) [4], *Locally Adaptive Random Transitions* (LART) [5], *Multiplex-Infomap* [6], *MultiGA* [7], *MultiMOGA* [8]

[1] D. LaSalle and G. Karypis, "Multi-threaded modularity based graph clustering using the multilevel paradigm", J. Parallel Distrib. Comput., 76:66–80, 2015.

[2] L. Tang, X. Wang, and H. Liu, "Uncovering groups via heterogeneous interaction analysis," in *Proc. ICDM*, 2009, pp. 503–512.

[3] M. Berlingerio, F. Pinelli, and F. Calabrese, "ABACUS: frequent pattern mining-based community discovery in multidimensional networks", Data Min. Knowl. Discov., 27(3):294– 320, 2013.

[4] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[5] Z. Kuncheva and G. Montana, "Community detection in multiplex networks using locally adaptive random walks," in *Proc. ASONAM*, 2015, pp. 1308–1315.

[6] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, "Identifying Modular Flows on Multilayer Networks Reveals HighlynOverlapping Organization in Interconnected Systems", Phys. Rev. X, 5, 011027, 2015.

[7] A. Amelio and C. Pizzuti, "A Cooperative Evolutionary Approach to Learn Communities in Multilayer Networks", In Proc. PSSN, pages 222–232, 2014.

[8] A. Amelio and C. Pizzuti, "Community detection in multidimensional networks", In Proc. ICTAI, pages 352–359, 2014.

Experimental evaluation

Assessment criteria

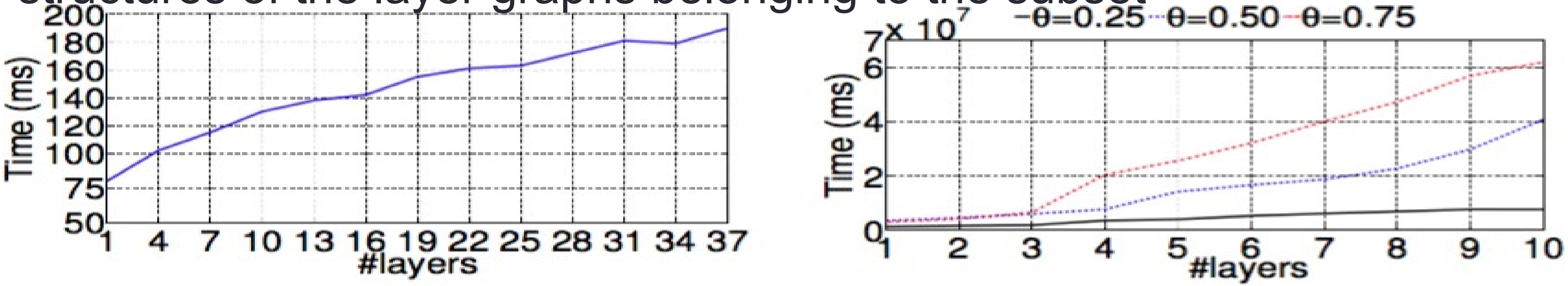
- **Redundancy** Berlingerio et al., ASONAM 2011
  - "redundant" connections, i.e., pairs of nodes connected through edges of different layers
- **Silhouette** – *extension to multilayer networks* Amelio, Tagarelli, CompleNet 2018
  - the distance computation terms are linearly combined over all layers
  - the distance between two nodes as one minus the Jaccard coefficient defined over the layer-specific sets of neighbors

Strehl, JMLR 2003; Dhillon, KDD 2004

- **NMI (2 definitions)**
  - vs. the solution obtained by Nerstrand on the flattened multilayer graph
  - vs. the layer-specific community structures

Efficiency of M-EMCD

- Layer graphs ordered by increasing size
- Several subsets by grouping the layer graphs according to their size order
  - For every subset, the ensemble corresponded to the community structures of the layer graphs belonging to the subset



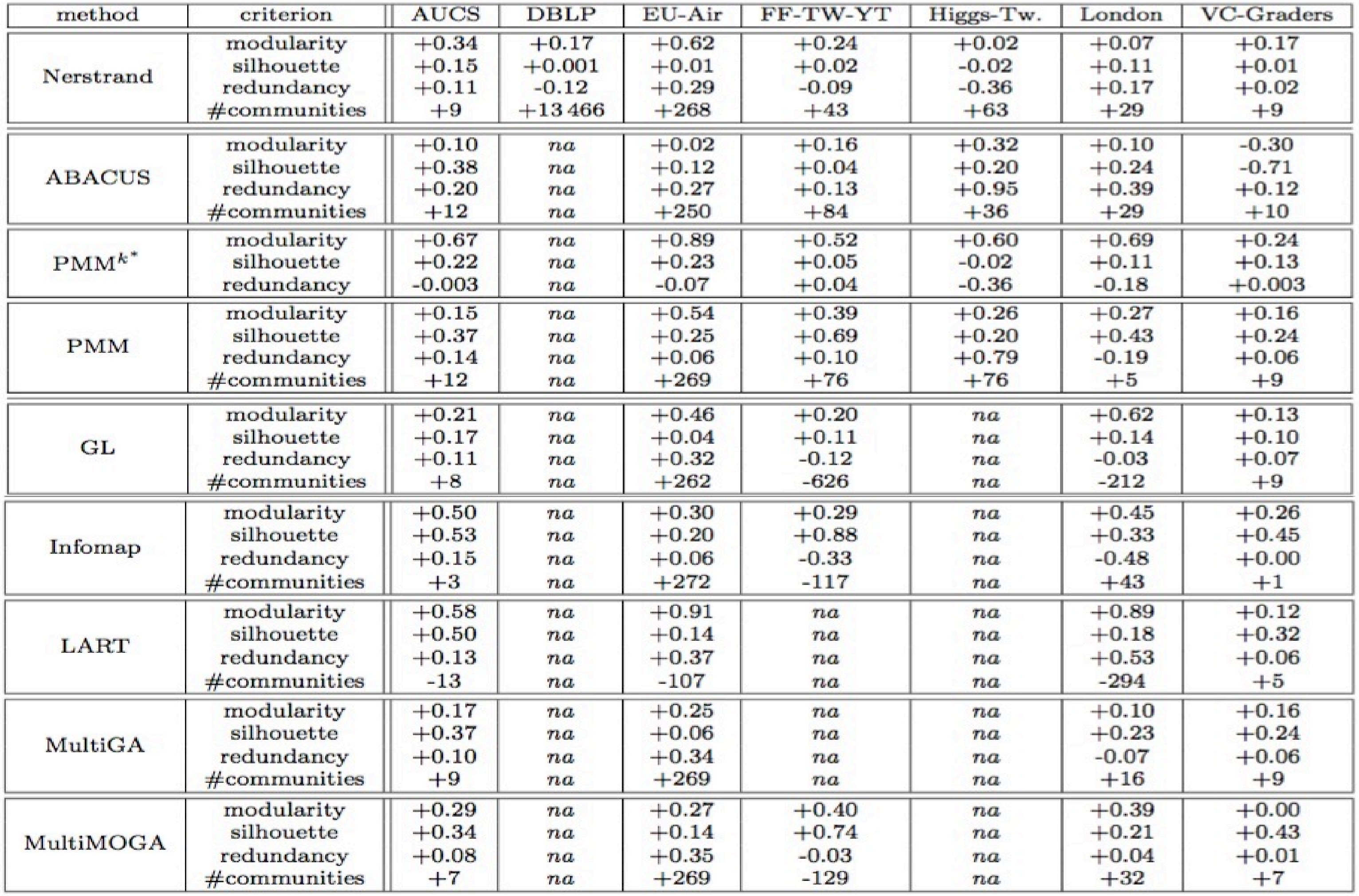
(a)

(b)

Time performance of M-EMCD on (a) EU-Air and (b) mLFR-1M

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

M-EMCD Gains vs. competing methods



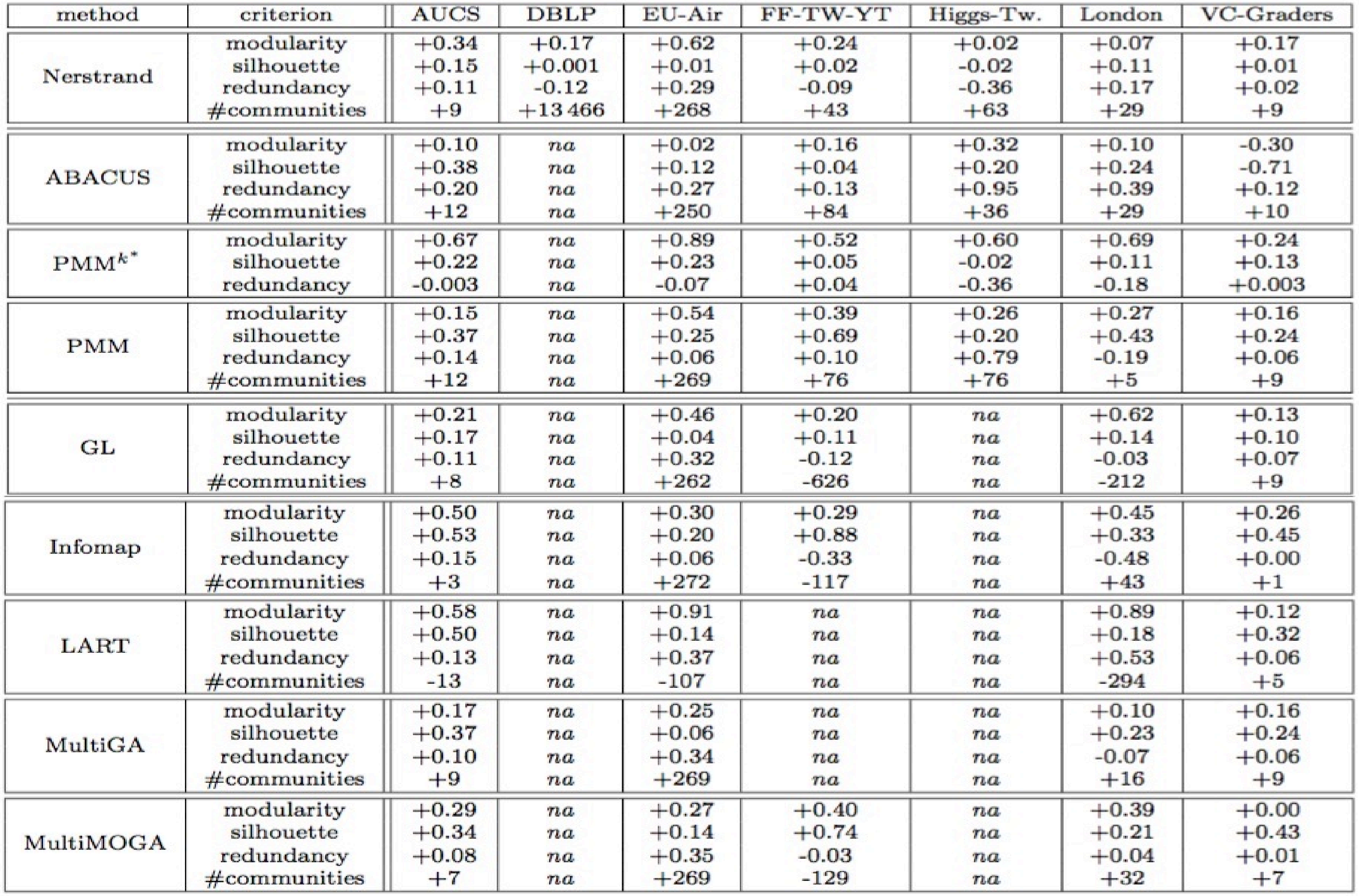
method	criterion	AUCS	DBLP	EU-Air	FF-TW-YT	Higgs-Tw.	London	VC-Graders
Nerstrand	modularity	+0.34	+0.17	+0.62	+0.24	+0.02	+0.07	+0.17
	silhouette	+0.15	+0.001	+0.01	+0.02	-0.02	+0.11	+0.01
	redundancy	+0.11	-0.12	+0.29	-0.09	-0.36	+0.17	+0.02
	#communities	+9	+13 466	+268	+43	+63	+29	+9
ABACUS	modularity	+0.10	*na*	+0.02	+0.16	+0.32	+0.10	-0.30
	silhouette	+0.38	*na*	+0.12	+0.04	+0.20	+0.24	-0.71
	redundancy	+0.20	*na*	+0.27	+0.13	+0.95	+0.39	+0.12
	#communities	+12	*na*	+250	+84	+36	+29	+10
PMM$^{k^*}$	modularity	+0.67	*na*	+0.89	+0.52	+0.60	+0.69	+0.24
	silhouette	+0.22	*na*	+0.23	+0.05	-0.02	+0.11	+0.13
	redundancy	-0.003	*na*	-0.07	+0.04	-0.36	-0.18	+0.003
PMM	modularity	+0.15	*na*	+0.54	+0.39	+0.26	+0.27	+0.16
	silhouette	+0.37	*na*	+0.25	+0.69	+0.20	+0.43	+0.24
	redundancy	+0.14	*na*	+0.06	+0.10	+0.79	-0.19	+0.06
	#communities	+12	*na*	+269	+76	+76	+5	+9
GL	modularity	+0.21	*na*	+0.46	+0.20	*na*	+0.62	+0.13
	silhouette	+0.17	*na*	+0.04	+0.11	*na*	+0.14	+0.10
	redundancy	+0.11	*na*	+0.32	-0.12	*na*	-0.03	+0.07
	#communities	+8	*na*	+262	-626	*na*	-212	+9
Infomap	modularity	+0.50	*na*	+0.30	+0.29	*na*	+0.45	+0.26
	silhouette	+0.53	*na*	+0.20	+0.88	*na*	+0.33	+0.45
	redundancy	+0.15	*na*	+0.06	-0.33	*na*	-0.48	+0.00
	#communities	+3	*na*	+272	-117	*na*	+43	+1
LART	modularity	+0.58	*na*	+0.91	*na*	*na*	+0.89	+0.12
	silhouette	+0.50	*na*	+0.14	*na*	*na*	+0.18	+0.32
	redundancy	+0.13	*na*	+0.37	*na*	*na*	+0.53	+0.06
	#communities	-13	*na*	-107	*na*	*na*	-294	+5
MultiGA	modularity	+0.17	*na*	+0.25	*na*	*na*	+0.10	+0.16
	silhouette	+0.37	*na*	+0.06	*na*	*na*	+0.23	+0.24
	redundancy	+0.10	*na*	+0.34	*na*	*na*	-0.07	+0.06
	#communities	+9	*na*	+269	*na*	*na*	+16	+9
MultiMOGA	modularity	+0.29	*na*	+0.27	+0.40	*na*	+0.39	+0.00
	silhouette	+0.34	*na*	+0.14	+0.74	*na*	+0.21	+0.43
	redundancy	+0.08	*na*	+0.35	-0.03	*na*	+0.04	+0.01
	#communities	+7	*na*	+269	-129	*na*	+32	+7

Modularity avg. gains:

0.63 vs. LART,
0.60 vs. PMMk,
0.36 vs. Infomap,
0.32 vs. GL,
0.30 vs PMM,
0.27 vs. MultiMOGA,
0.23 vs. Nerstrand,
0.17 vs. MultiGA,
and 0.07 vs. ABACUS.

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

M-EMCD Gains vs. competing methods



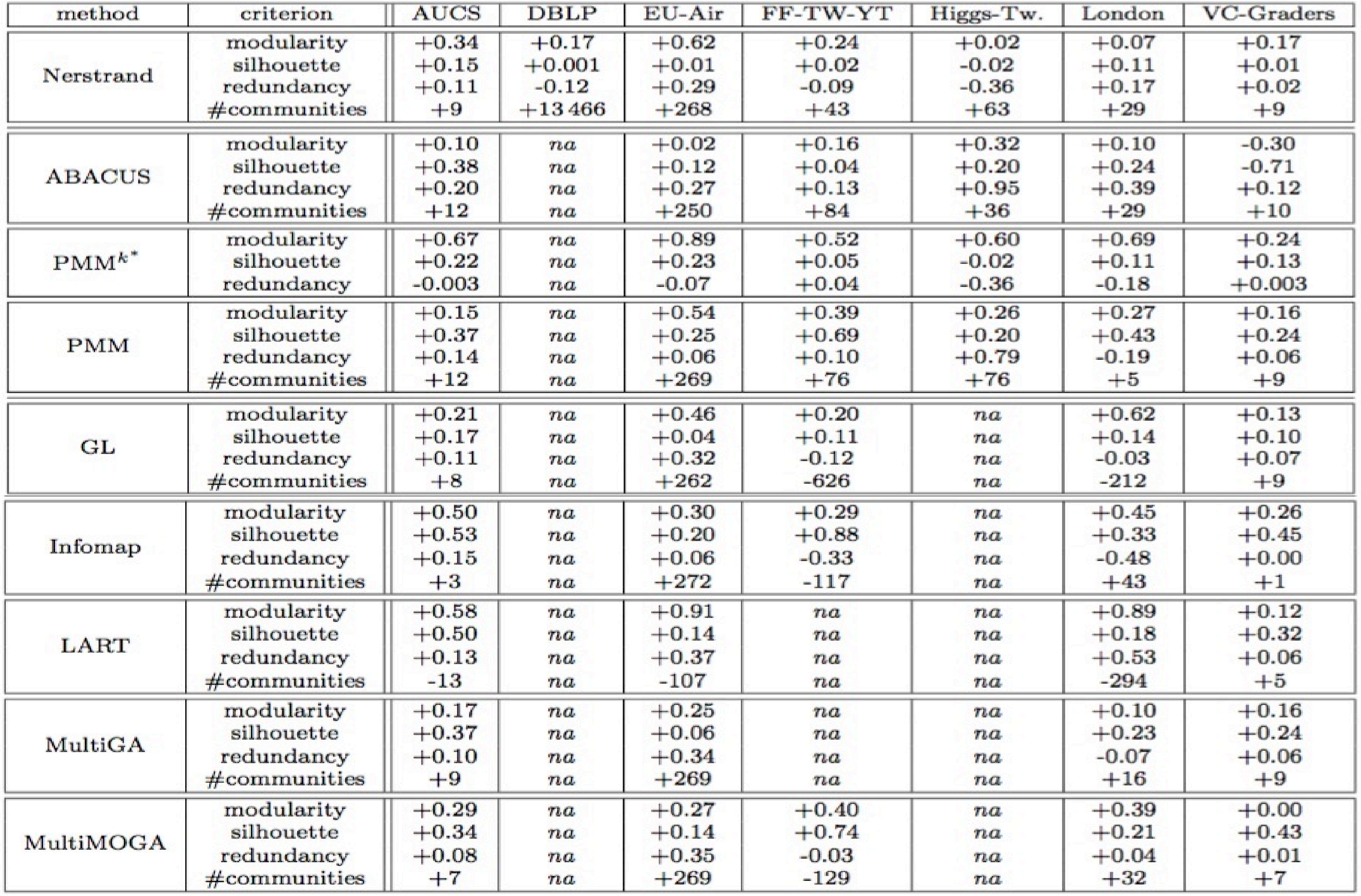
method	criterion	AUCS	DBLP	EU-Air	FF-TW-YT	Higgs-Tw.	London	VC-Graders
Nerstrand	modularity	+0.34	+0.17	+0.62	+0.24	+0.02	+0.07	+0.17
	silhouette	+0.15	+0.001	+0.01	+0.02	-0.02	+0.11	+0.01
	redundancy	+0.11	-0.12	+0.29	-0.09	-0.36	+0.17	+0.02
	#communities	+9	+13 466	+268	+43	+63	+29	+9
ABACUS	modularity	+0.10	*na*	+0.02	+0.16	+0.32	+0.10	-0.30
	silhouette	+0.38	*na*	+0.12	+0.04	+0.20	+0.24	-0.71
	redundancy	+0.20	*na*	+0.27	+0.13	+0.95	+0.39	+0.12
	#communities	+12	*na*	+250	+84	+36	+29	+10
PMM$^{k^*}$	modularity	+0.67	*na*	+0.89	+0.52	+0.60	+0.69	+0.24
	silhouette	+0.22	*na*	+0.23	+0.05	-0.02	+0.11	+0.13
	redundancy	-0.003	*na*	-0.07	+0.04	-0.36	-0.18	+0.003
PMM	modularity	+0.15	*na*	+0.54	+0.39	+0.26	+0.27	+0.16
	silhouette	+0.37	*na*	+0.25	+0.69	+0.20	+0.43	+0.24
	redundancy	+0.14	*na*	+0.06	+0.10	+0.79	-0.19	+0.06
	#communities	+12	*na*	+269	+76	+76	+5	+9
GL	modularity	+0.21	*na*	+0.46	+0.20	*na*	+0.62	+0.13
	silhouette	+0.17	*na*	+0.04	+0.11	*na*	+0.14	+0.10
	redundancy	+0.11	*na*	+0.32	-0.12	*na*	-0.03	+0.07
	#communities	+8	*na*	+262	-626	*na*	-212	+9
Infomap	modularity	+0.50	*na*	+0.30	+0.29	*na*	+0.45	+0.26
	silhouette	+0.53	*na*	+0.20	+0.88	*na*	+0.33	+0.45
	redundancy	+0.15	*na*	+0.06	-0.33	*na*	-0.48	+0.00
	#communities	+3	*na*	+272	-117	*na*	+43	+1
LART	modularity	+0.58	*na*	+0.91	*na*	*na*	+0.89	+0.12
	silhouette	+0.50	*na*	+0.14	*na*	*na*	+0.18	+0.32
	redundancy	+0.13	*na*	+0.37	*na*	*na*	+0.53	+0.06
	#communities	-13	*na*	-107	*na*	*na*	-294	+5
MultiGA	modularity	+0.17	*na*	+0.25	*na*	*na*	+0.10	+0.16
	silhouette	+0.37	*na*	+0.06	*na*	*na*	+0.23	+0.24
	redundancy	+0.10	*na*	+0.34	*na*	*na*	-0.07	+0.06
	#communities	+9	*na*	+269	*na*	*na*	+16	+9
MultiMOGA	modularity	+0.29	*na*	+0.27	+0.40	*na*	+0.39	+0.00
	silhouette	+0.34	*na*	+0.14	+0.74	*na*	+0.21	+0.43
	redundancy	+0.08	*na*	+0.35	-0.03	*na*	+0.04	+0.01
	#communities	+7	*na*	+269	-129	*na*	+32	+7

Silhouette avg. gains:

0.48 vs. Multiplex-Infomap,
0.37 vs. MultiMOGA,
0.36 vs. PMM,
0.29 vs. LART,
0.23 vs. MultiGA,
0.12 vs. PMMk,
0.11 vs. GL,
0.05 vs. ABACUS,
and 0.04 vs. Nerstrand.

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

M-EMCD Gains vs. competing methods



method	criterion	AUCS	DBLP	EU-Air	FF-TW-YT	Higgs-Tw.	London	VC-Graders
Nerstrand	modularity	+0.34	+0.17	+0.62	+0.24	+0.02	+0.07	+0.17
	silhouette	+0.15	+0.001	+0.01	+0.02	-0.02	+0.11	+0.01
	redundancy	+0.11	-0.12	+0.29	-0.09	-0.36	+0.17	+0.02
	#communities	+9	+13 466	+268	+43	+63	+29	+9
ABACUS	modularity	+0.10	*na*	+0.02	+0.16	+0.32	+0.10	-0.30
	silhouette	+0.38	*na*	+0.12	+0.04	+0.20	+0.24	-0.71
	redundancy	+0.20	*na*	+0.27	+0.13	+0.95	+0.39	+0.12
	#communities	+12	*na*	+250	+84	+36	+29	+10
PMM$^{k^*}$	modularity	+0.67	*na*	+0.89	+0.52	+0.60	+0.69	+0.24
	silhouette	+0.22	*na*	+0.23	+0.05	-0.02	+0.11	+0.13
	redundancy	-0.003	*na*	-0.07	+0.04	-0.36	-0.18	+0.003
PMM	modularity	+0.15	*na*	+0.54	+0.39	+0.26	+0.27	+0.16
	silhouette	+0.37	*na*	+0.25	+0.69	+0.20	+0.43	+0.24
	redundancy	+0.14	*na*	+0.06	+0.10	+0.79	-0.19	+0.06
	#communities	+12	*na*	+269	+76	+76	+5	+9
GL	modularity	+0.21	*na*	+0.46	+0.20	*na*	+0.62	+0.13
	silhouette	+0.17	*na*	+0.04	+0.11	*na*	+0.14	+0.10
	redundancy	+0.11	*na*	+0.32	-0.12	*na*	-0.03	+0.07
	#communities	+8	*na*	+262	-626	*na*	-212	+9
Infomap	modularity	+0.50	*na*	+0.30	+0.29	*na*	+0.45	+0.26
	silhouette	+0.53	*na*	+0.20	+0.88	*na*	+0.33	+0.45
	redundancy	+0.15	*na*	+0.06	-0.33	*na*	-0.48	+0.00
	#communities	+3	*na*	+272	-117	*na*	+43	+1
LART	modularity	+0.58	*na*	+0.91	*na*	*na*	+0.89	+0.12
	silhouette	+0.50	*na*	+0.14	*na*	*na*	+0.18	+0.32
	redundancy	+0.13	*na*	+0.37	*na*	*na*	+0.53	+0.06
	#communities	-13	*na*	-107	*na*	*na*	-294	+5
MultiGA	modularity	+0.17	*na*	+0.25	*na*	*na*	+0.10	+0.16
	silhouette	+0.37	*na*	+0.06	*na*	*na*	+0.23	+0.24
	redundancy	+0.10	*na*	+0.34	*na*	*na*	-0.07	+0.06
	#communities	+9	*na*	+269	*na*	*na*	+16	+9
MultiMOGA	modularity	+0.29	*na*	+0.27	+0.40	*na*	+0.39	+0.00
	silhouette	+0.34	*na*	+0.14	+0.74	*na*	+0.21	+0.43
	redundancy	+0.08	*na*	+0.35	-0.03	*na*	+0.04	+0.01
	#communities	+7	*na*	+269	-129	*na*	+32	+7

Global redundancy:

- higher than ABACUS and LART
- lower than the other methods

Coupling modularity, silhouette and redundancy results:

M-EMCD can utilize less information from the various layers than other methods to obtain higher quality consensus community structures

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Summary of findings

- M-EMCD outperforms CC-EMCD and C-EMCD methods
  - in terms of modularity as well as silhouette of community membership
  - redundancy comparable to C-EMCD
- M-EMCD is relatively robust against
  - presence of disconnected components in a multilayer graph
    - small number of singleton communities in the consensus solutions
  - perturbations in the input ensemble (size of clusterings)
- M-EMCD scales well with the size of a multilayer network
  - Linear cost in the number of edges
- M-EMCD outperforms competing methods
  - in terms of modularity as well as silhouette of community membership
  - tends to use less multilayer information

A. Tagarelli et al. *Ensemble-based community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Parameter-free graph pruning Enhanced M-EMCD

Limitations of EMCD

- The co-association matrix filtering relies on a user-specified threshold (θ)
  - Guessing best θ is network-dependent
  - θ-based pruning discards properties related to node distributions

- "Static" community membership of nodes during modularity optimization in M-EMCD

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Solutions for enhanced EMCD

- Parameter-free identification of consensus clusters based on *generative models* for graph pruning
  - Goal: filter out noisy edges from the EMCD co-association matrix
  - Key aspects:
    - No requirements for any user-specified parameter
    - Edge-removal decision is taken according to statistical significance (based on node degree/strength distributions)

- 3-stage M-EMCD iterative scheme
  - Intra-community connectivity refinement
  - Community partitioning
  - Inter-community connectivity refinement with relocation of nodes to neighboring communities

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Generative models for graph pruning (1/2)

General scheme

Given a weighted undirected graph:

1. Define a *null model* based on node distribution properties
2. Compute a *p-value* for every edge
   - to determine the statistical significance of properties assigned to edges from a given distribution
3. Filter out all edges having p-value above a chosen significance level
   - *i.e.*, keep all edges that are least likely to have occurred due to random chance

Gemmetto, V., Cardillo, A., Garlaschelli, D.: Irreducible network backbones: unbiased graph filtering via maximum entropy. *arXiv* (June 2017)

Dianati, N.: Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical Review E* 93, 012304 (2016)

Radicchi, F., Ramasco, J.J., Fortunato, S.: Information filtering in complex weighted networks. *Physical Review E* 83, 046101 (2011)

Parameter-free co-association filtering (1/2)

- θ-based pruning
  - too low values may lead to few, large communities, while too high values may lead to many, small communities
  - Iterative search for the best-performing θ does not scale

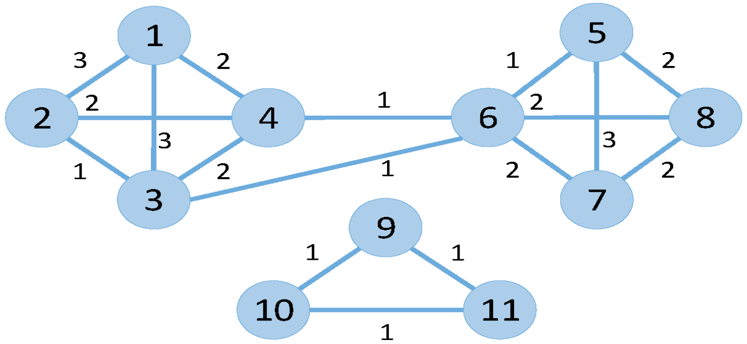
Example:

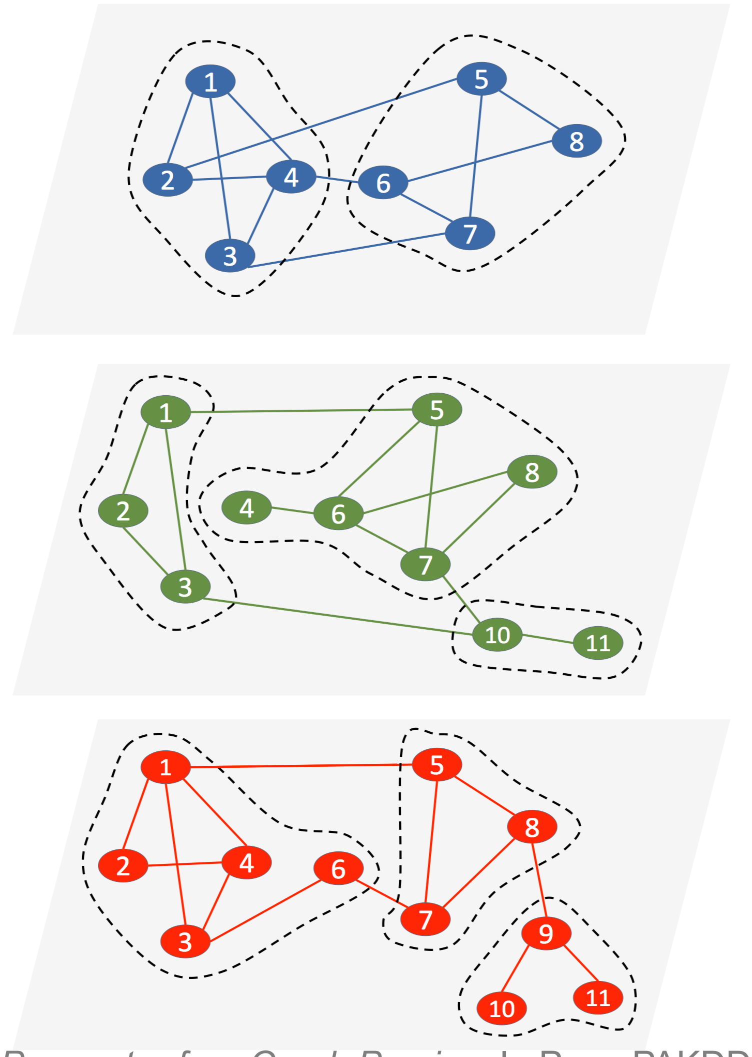
- θ δ 1/3

  C_1={1,..,8}, C_2={9,10,11}
- 1/3 < θ δ 2/3

  C_1={1,..,4}, C_2={5,..,8}, C_3={9}, C_4={10}, C_5={11}
- θ > 2/3

  C_1={1, 2, 3}, C_2={5,7}, plus 5 singleton communities





D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Parameter-free co-association filtering (2/2)

- Can we account for the multilayer network topology to evaluate the significance of the co-associations?
  - low value of co-association: may correspond to node relations pertinent to some of the layers ➔ *meaningful?*
  - high value of co-association: may correspond to the linkage of high-degree nodes co-occurring in the same community in many layers ➔ *superfluous?*

- **Idea**: evaluate a generative model for graph pruning over the weighted co-association graph

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Parameter-free co-association filtering

Co-association hypothesis testing

- Let WGP be a statistical inference method whose generative null model is param. w.r.t. node degree/strength distributions in the co-association graph
- Co-association hypothesis testing based on WGP:
  - Null hypothesis (H_0): observed edge weight generated by chance
  - p-value: prob. that the null model produces a weight ≥ the observed weight
- If the p-value is lower than α, then H_0 can be rejected

  ➔ the co-association is statist. meaningful

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Enhanced M-EMCD (1/2)

1. Incorporates parameter-free pruning of co-associations
   - Prior to generation of lower-bound consensus (i.e., CC-EMCD solution)

2. Dynamically adjusts memberships during the consensus optimization – 3 stages

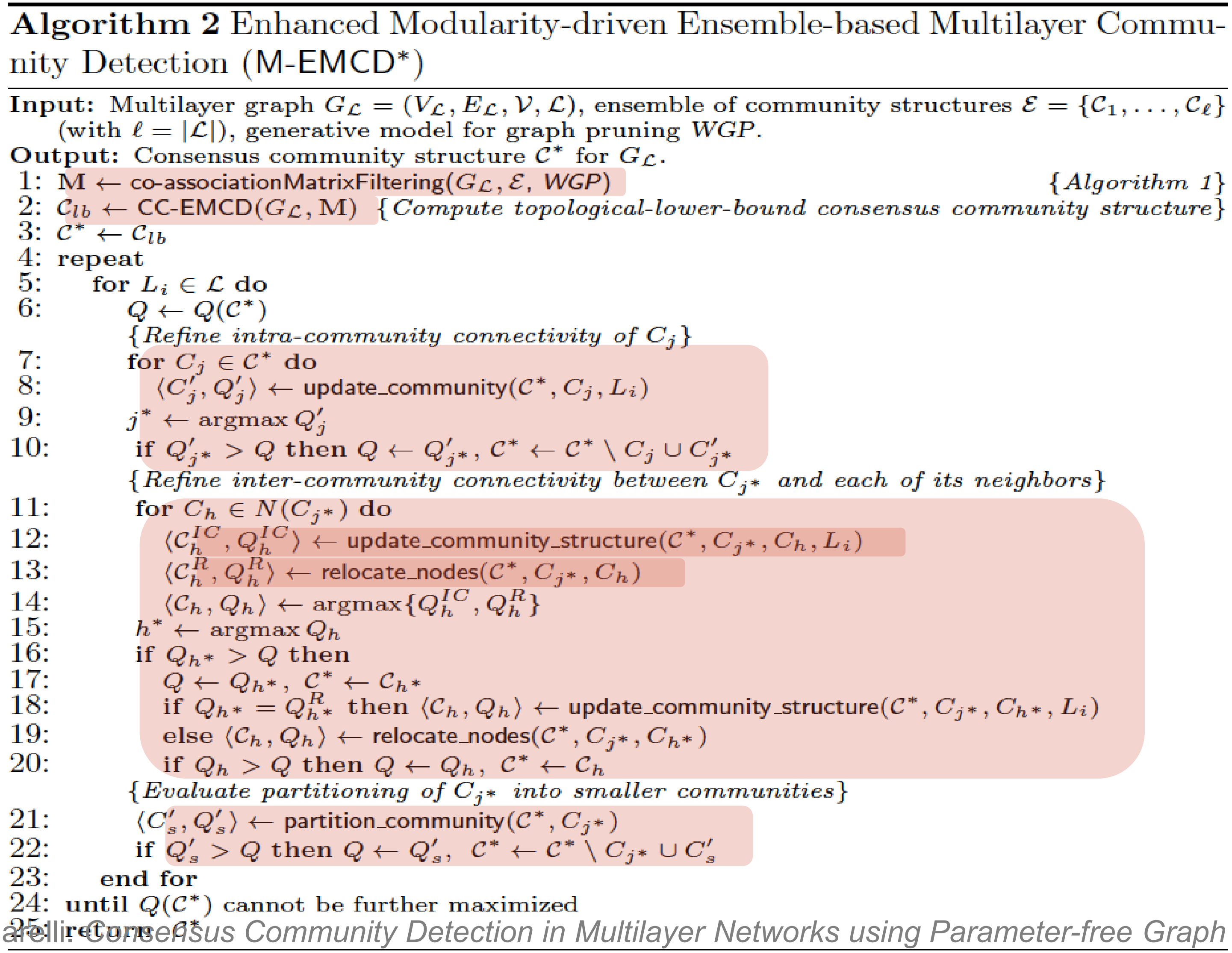
   (i) refinement of connectivity internal to a selected community

   (ii) refinement of connectivity between the community and its neighbors, with relocation of nodes

   (iii) partitioning of the community

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

Enhanced M-EMCD: **M-EMCD*** (2/2)



Algorithm 2 Enhanced Modularity-driven Ensemble-based Multilayer Community Detection (M-EMCD*)

```
Input: Multilayer graph $G_{\mathcal{L}} = (V_{\mathcal{L}}, E_{\mathcal{L}}, \mathcal{V}, \mathcal{L})$, ensemble of community structures $\mathcal{E} = \{\mathcal{C}_1, \ldots, \mathcal{C}_\ell\}$
    (with $\ell = |\mathcal{L}|$), generative model for graph pruning $WGP$.
Output: Consensus community structure $\mathcal{C}^*$ for $G_{\mathcal{L}}$.
M ← co-associationMatrixFiltering($G_{\mathcal{L}}, \mathcal{E}$, $WGP$)    {Algorithm 1}
$\mathcal{C}_{lb}$ ← CC-EMCD($G_{\mathcal{L}}$, M)  {Compute topological-lower-bound consensus community structure}
$\mathcal{C}^* \leftarrow \mathcal{C}_{lb}$
repeat
    for $L_i \in \mathcal{L}$ do
        $Q \leftarrow Q(\mathcal{C}^*)$
        {Refine intra-community connectivity of $C_j$}
        for $C_j \in \mathcal{C}^*$ do
            $\langle C'_j, Q'_j \rangle$ ← update_community($\mathcal{C}^*, C_j, L_i$)
        $j^* \leftarrow \mathrm{argmax}\, Q'_j$
        if $Q'_{j^*} > Q$ then $Q \leftarrow Q'_{j^*}$, $\mathcal{C}^* \leftarrow \mathcal{C}^* \setminus C_j \cup C'_{j^*}$
        {Refine inter-community connectivity between $C_{j^*}$ and each of its neighbors}
        for $C_h \in N(C_{j^*})$ do
            $\langle \mathcal{C}_h^{IC}, Q_h^{IC} \rangle$ ← update_community_structure($\mathcal{C}^*, C_{j^*}, C_h, L_i$)
            $\langle \mathcal{C}_h^{R}, Q_h^{R} \rangle$ ← relocate_nodes($\mathcal{C}^*, C_{j^*}, C_h$)
            $\langle \mathcal{C}_h, Q_h \rangle \leftarrow \mathrm{argmax}\{Q_h^{IC}, Q_h^{R}\}$
        $h^* \leftarrow \mathrm{argmax}\, Q_h$
        if $Q_{h^*} > Q$ then
            $Q \leftarrow Q_{h^*}$, $\mathcal{C}^* \leftarrow \mathcal{C}_{h^*}$
            if $Q_{h^*} = Q^R_{h^*}$ then $\langle \mathcal{C}_h, Q_h \rangle$ ← update_community_structure($\mathcal{C}^*, C_{j^*}, C_{h^*}, L_i$)
            else $\langle \mathcal{C}_h, Q_h \rangle$ ← relocate_nodes($\mathcal{C}^*, C_{j^*}, C_{h^*}$)
            if $Q_h > Q$ then $Q \leftarrow Q_h$, $\mathcal{C}^* \leftarrow \mathcal{C}_h$
        {Evaluate partitioning of $C_{j^*}$ into smaller communities}
        $\langle C'_s, Q'_s \rangle$ ← partition_community($\mathcal{C}^*, C_{j^*}$)
        if $Q'_s > Q$ then $Q \leftarrow Q'_s$, $\mathcal{C}^* \leftarrow \mathcal{C}^* \setminus C_{j^*} \cup C'_s$
    end for
until $Q(\mathcal{C}^*)$ cannot be further maximized
return $\mathcal{C}^*$
```

D. Mandaglio, A. Amelio, A. Tagarelli: *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

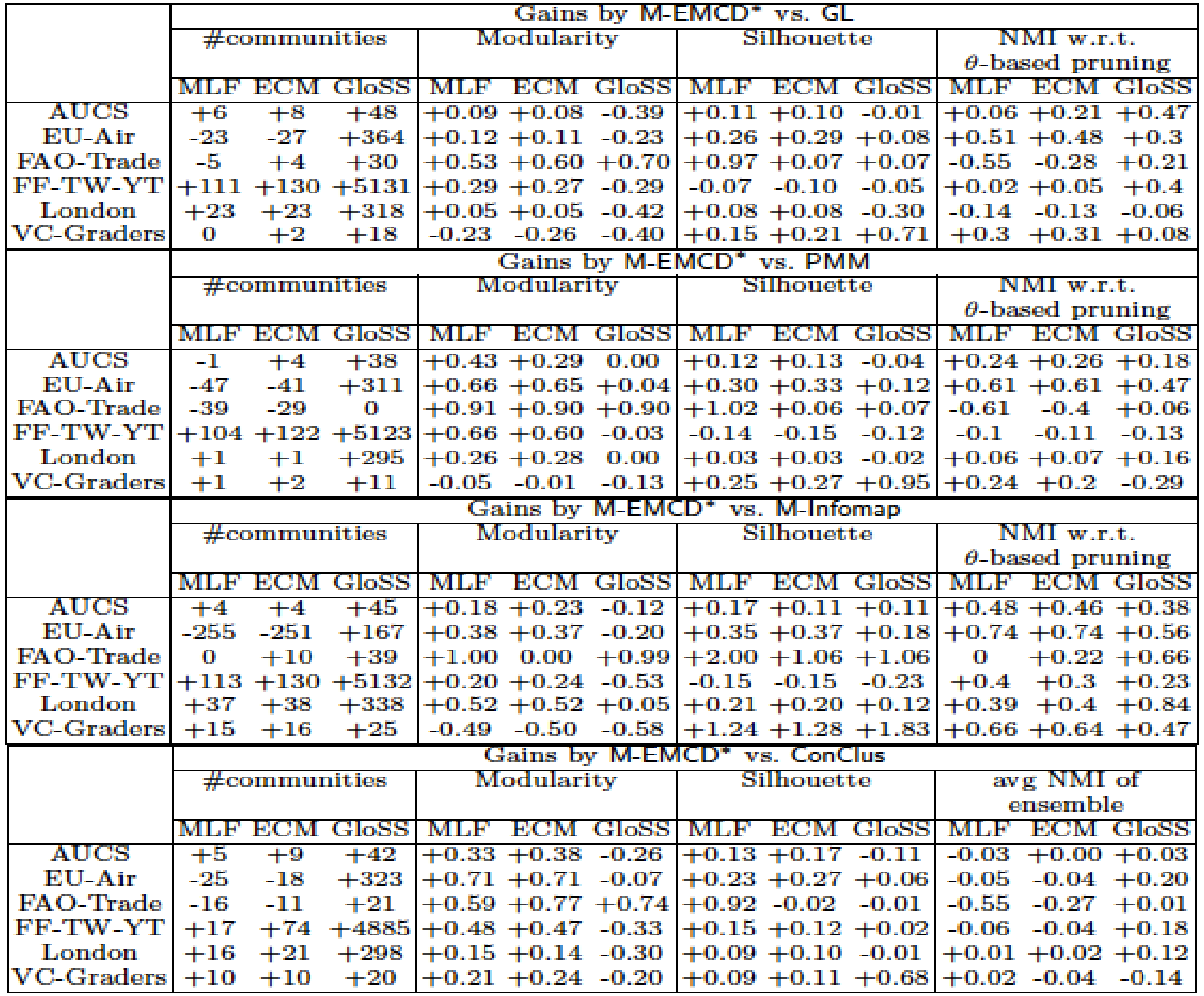
Enhanced M-EMCD Gains vs. competing methods

Direct:

- **GL** (Mucha et al. 2010)
- **M-Infomap** (De Domenico et al. 2015)

Aggregate:

- **PMM** (Tang et al. 2009)
  - Varying no. of communities
- Consensus-clustering (**ConClus**) (Lancichinetti & Fortunato, 2012)
  - Nerstrand for generating the clustering solutions
  - n_p set to the no. of layers
  - Selection of θ corresponding to the largest NMI w.r.t. the initial clusterings



	Gains by M-EMCD* vs. GL											
	#communities			Modularity			Silhouette			NMI w.r.t. θ-based pruning		
	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS
AUCS	+6	+8	+48	+0.09	+0.08	-0.39	+0.11	+0.10	-0.01	+0.06	+0.21	+0.47
EU-Air	-23	-27	+364	+0.12	+0.11	-0.23	+0.26	+0.29	+0.08	+0.51	+0.48	+0.3
FAO-Trade	-5	+4	+30	+0.53	+0.60	+0.70	+0.97	+0.07	+0.07	-0.55	-0.28	+0.21
FF-TW-YT	+111	+130	+5131	+0.29	+0.27	-0.29	-0.07	-0.10	-0.05	+0.02	+0.05	+0.4
London	+23	+23	+318	+0.05	+0.05	-0.42	+0.08	+0.08	-0.30	-0.14	-0.13	-0.06
VC-Graders	0	+2	+18	-0.23	-0.26	-0.40	+0.15	+0.21	+0.71	+0.3	+0.31	+0.08

	Gains by M-EMCD* vs. PMM											
	#communities			Modularity			Silhouette			NMI w.r.t. θ-based pruning		
	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS
AUCS	-1	+4	+38	+0.43	+0.29	0.00	+0.12	+0.13	-0.04	+0.24	+0.26	+0.18
EU-Air	-47	-41	+311	+0.66	+0.65	+0.04	+0.30	+0.33	+0.12	+0.61	+0.61	+0.47
FAO-Trade	-39	-29	0	+0.91	+0.90	+0.90	+1.02	+0.06	+0.07	-0.61	-0.4	+0.06
FF-TW-YT	+104	+122	+5123	+0.66	+0.60	-0.03	-0.14	-0.15	-0.12	-0.1	-0.11	-0.13
London	+1	+1	+295	+0.26	+0.28	0.00	+0.03	+0.03	-0.02	+0.06	+0.07	+0.16
VC-Graders	+1	+2	+11	-0.05	-0.01	-0.13	+0.25	+0.27	+0.95	+0.24	+0.2	-0.29

	Gains by M-EMCD* vs. M-Infomap											
	#communities			Modularity			Silhouette			NMI w.r.t. θ-based pruning		
	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS
AUCS	+4	+4	+45	+0.18	+0.23	-0.12	+0.17	+0.11	+0.11	+0.48	+0.46	+0.38
EU-Air	-255	-251	+167	+0.38	+0.37	-0.20	+0.35	+0.37	+0.18	+0.74	+0.74	+0.56
FAO-Trade	0	+10	+39	+1.00	0.00	+0.99	+2.00	+1.06	+1.06	0	+0.22	+0.66
FF-TW-YT	+113	+130	+5132	+0.20	+0.24	-0.53	-0.15	-0.15	-0.23	+0.4	+0.3	+0.23
London	+37	+38	+338	+0.52	+0.52	+0.05	+0.21	+0.20	+0.12	+0.39	+0.4	+0.84
VC-Graders	+15	+16	+25	-0.49	-0.50	-0.58	+1.24	+1.28	+1.83	+0.66	+0.64	+0.47

	Gains by M-EMCD* vs. ConClus											
	#communities			Modularity			Silhouette			avg NMI of ensemble		
	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS	MLF	ECM	GloSS
AUCS	+5	+9	+42	+0.33	+0.38	-0.26	+0.13	+0.17	-0.11	-0.03	+0.00	+0.03
EU-Air	-25	-18	+323	+0.71	+0.71	-0.07	+0.23	+0.27	+0.06	-0.05	-0.04	+0.20
FAO-Trade	-16	-11	+21	+0.59	+0.77	+0.74	+0.92	-0.02	-0.01	-0.55	-0.27	+0.01
FF-TW-YT	+17	+74	+4885	+0.48	+0.47	-0.33	+0.15	+0.12	+0.02	-0.06	-0.04	+0.18
London	+16	+21	+298	+0.15	+0.14	-0.30	+0.09	+0.10	-0.01	+0.01	+0.02	+0.12
VC-Graders	+10	+10	+20	+0.21	+0.24	-0.20	+0.09	+0.11	+0.68	+0.02	-0.04	-0.14

D. Mandaglio, A. Amelio, A. Tagarelli. *Consensus Community Detection in Multilayer Networks using Parameter-free Graph Pruning*. In Proc. PAKDD 2018

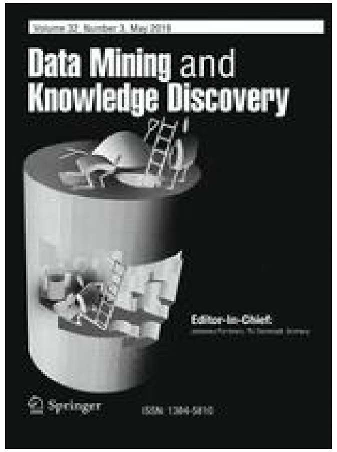
NEIGHBORHOOD AS CLUSTER IN A COMPLEX NETWORK SYSTEM

Node-centric (or local) multilayer community detection

Main references for this part

The 2017 European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases

ECML PKDD 2017



R. Interdonato, A. Tagarelli, D. Ienco, A. Sallaberry, P. Poncelet
Local community detection in multilayer networks.
Data Min. Knowl. Discov. 31(5): 1444-1479 (2017)

R. Interdonato, A. Tagarelli
Personalized Recommendation of Points-of-Interest based on Multilayer Local Community Detection
In *Proc. SocInfo 2017*

SocInfo 2017

Local community detection (1/2)

Conventional **Community Detection**: a global optimization problem -- it requires knowledge on the whole network structure

LCD is a different problem: finding a relatively expanded neighborhood of a single node which forms a densely connected, small subgraph

- e.g., personalized network of social contacts of interest to a single (or few) user only

Minimal requirements of memory-footprint

Key-enabling for dynamic network analysis

Useful to cope with privacy and access restriction issues

GOAL: Given limited information about the network, to identify a *community* which is centered on one (or few) seed nodes

Local community detection (2/2)

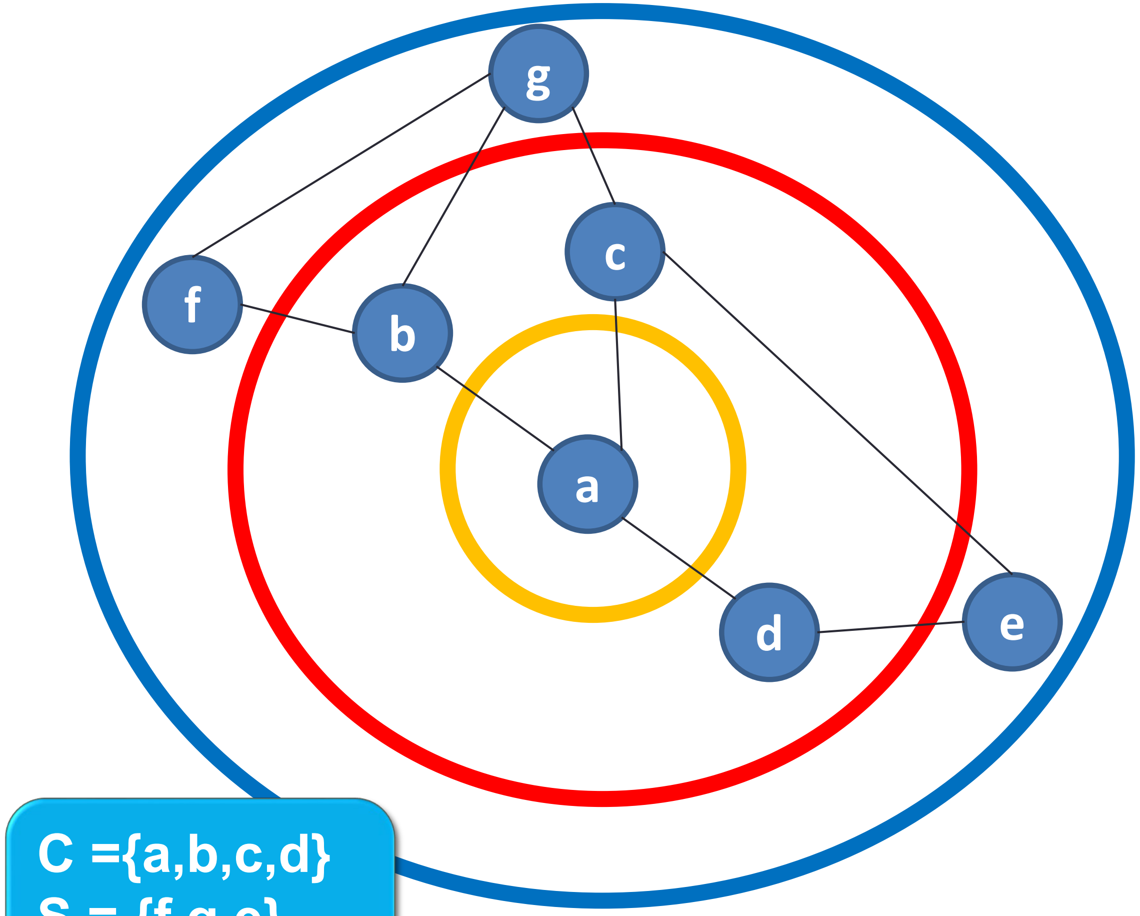
General approach (Clauset et al. 2005, Chen et al. 2009, Branting 2012, Fagnan et al. 2014)**:**

- accounting for the relative ratio of internal edges and external edges
- penalizing candidates in proportion to the amount of links to non-community nodes

Community (*C*): the local community under construction

Shell set (*S*): the set of neighbors of nodes in *C* that do not belong to *C*

Boundary set (*B*): subset of *C* comprised of nodes having neighbors in *S*



C ={a,b,c,d}
S = {f,g,e}
B = {b,c,d}

The ML-LCD problem

Given a multilayer graph $G_L = (V_L, E_L, V, L)$ with set of nodes V, and a seed node $v_0 \in V$, find a subgraph $G_L^{v_o} \subseteq G_L$ that contains v_0 and maximizes the multilayer local community function LC:

$$G_L^{v_o} = \mathrm{argmax}_{G \subseteq G_L} \frac{LC^{int}(G)}{LC^{ext}(G)}$$

- $LC^{int}(G)$: function proportional to the density of links among the nodes within G
- $LC^{ext}(G)$: function proportional to the density of links between the nodes within G and nodes outside G

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Layer-weighting-based local community functions

- Exploits layer relevance weighting scheme

- Simple linear combination over layers

$$LC^{int}(C) = \frac{1}{|C|} \sum_{v \in C} \sum_{L_i \in \mathcal{L}} \omega_i |E_i^C(v)|$$

$$LC^{ext}(C) = \frac{1}{|B|} \sum_{u \in B} \sum_{L_i \in \mathcal{L}} \omega_i |E_i^B(v)|$$

ML-LCD-*lw*

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Within-layer similarity-based local community functions

- Relies on a notion of **similarity of nodes**
- Any similarity measure that can express the topological affinity of two nodes in a graph
  - Jaccard sim: $sim_i(u,v) = \frac{|N_i(u) \cap N_i(v)|}{|N_i(u) \cup N_i(v)|}$
  - cosine sim
  - 3-clique based measures
  - (Node embeddings)
- Similarity between any two nodes is determined by **focusing on each layer at a time**

$$LC^{int}(C) = \frac{1}{|C|} \sum_{v \in C} \sum_{L_i \in \mathcal{L}} \sum_{\substack{(u,v) \in E_i^C \\ \wedge u \in C}} sim_i(u,v)$$

$$LC^{ext}(C) = \frac{1}{|B|} \sum_{v \in B} \sum_{L_i \in \mathcal{L}} \sum_{\substack{(u,v) \in E_i^B \\ \wedge u \in S}} sim_i(u,v)$$

ML-LCD-*wlsim*

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Cross-layer similarity-based local community functions

- *Topologyless* approach
  - Similarities are computed **among all nodes in set B and set S**, regardless of node relations
- Captures ties between two nodes that can still be significant even **without the presence of an explicit edge in the network**

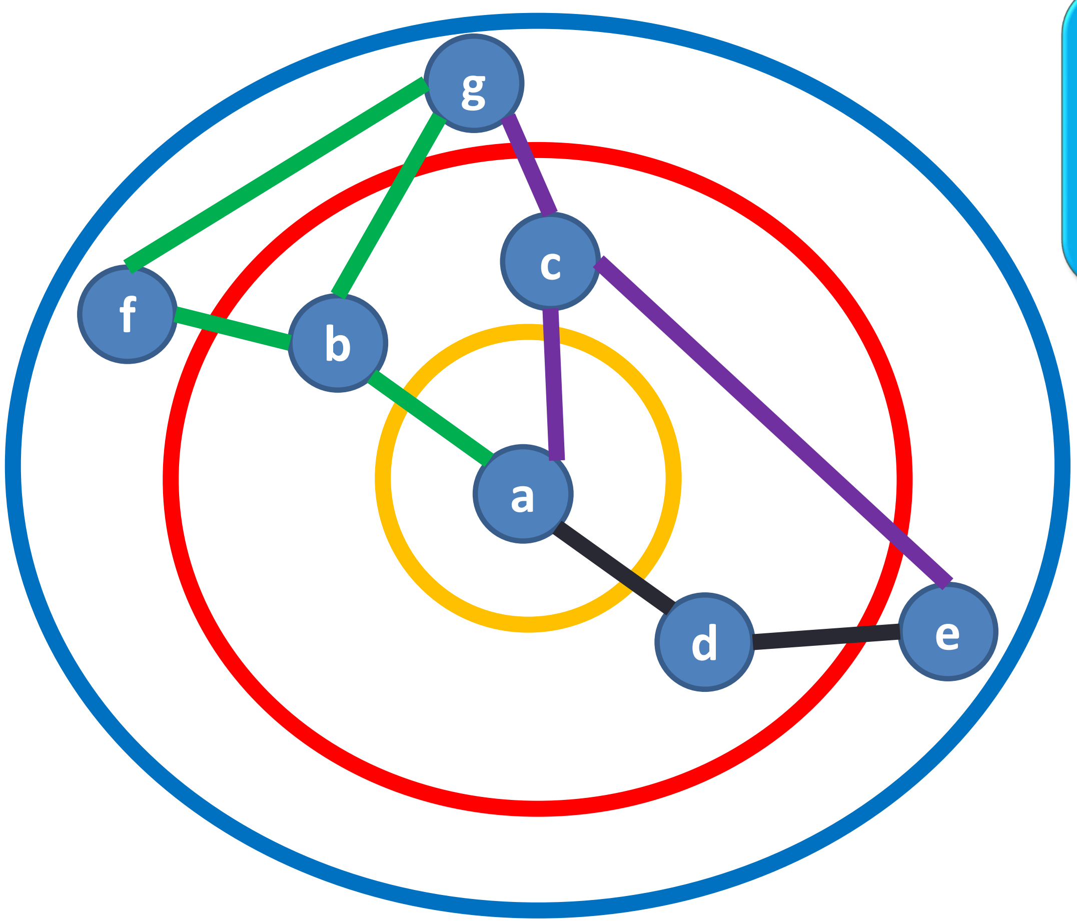
$$LC^{int}(C) = \frac{1}{|C|} \sum_{u,v \in C} \sum_{\substack{L_i, L_j \in \mathcal{L} \\ \wedge u \in V_i, v \in V_j}} sim_{i,j}(u,v)$$

$$LC^{ext}(C) = \frac{1}{|B|} \sum_{v \in B, u \in S} \sum_{\substack{L_i, L_j \in \mathcal{L} \\ \wedge u \in V_i, v \in V_j}} sim_{i,j}(u,v)$$

ML-LCD-*clsim*

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Cross-layer similarity-based local community functions

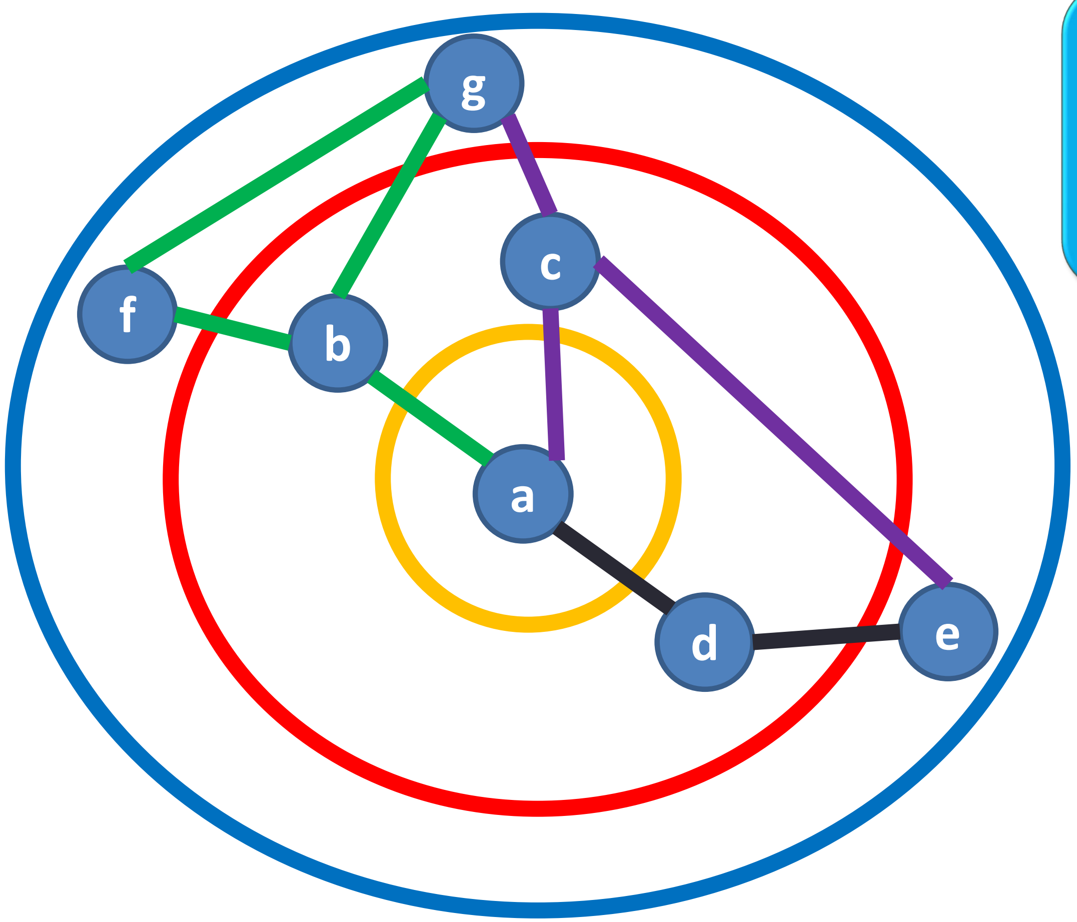


C = a,b,c,d
S = f,g,e
B = b,c,d

wl-sim:
sim(b,f)
sim(b,g)
sim(c,g)
sim(c,e)
sim(d,e)

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

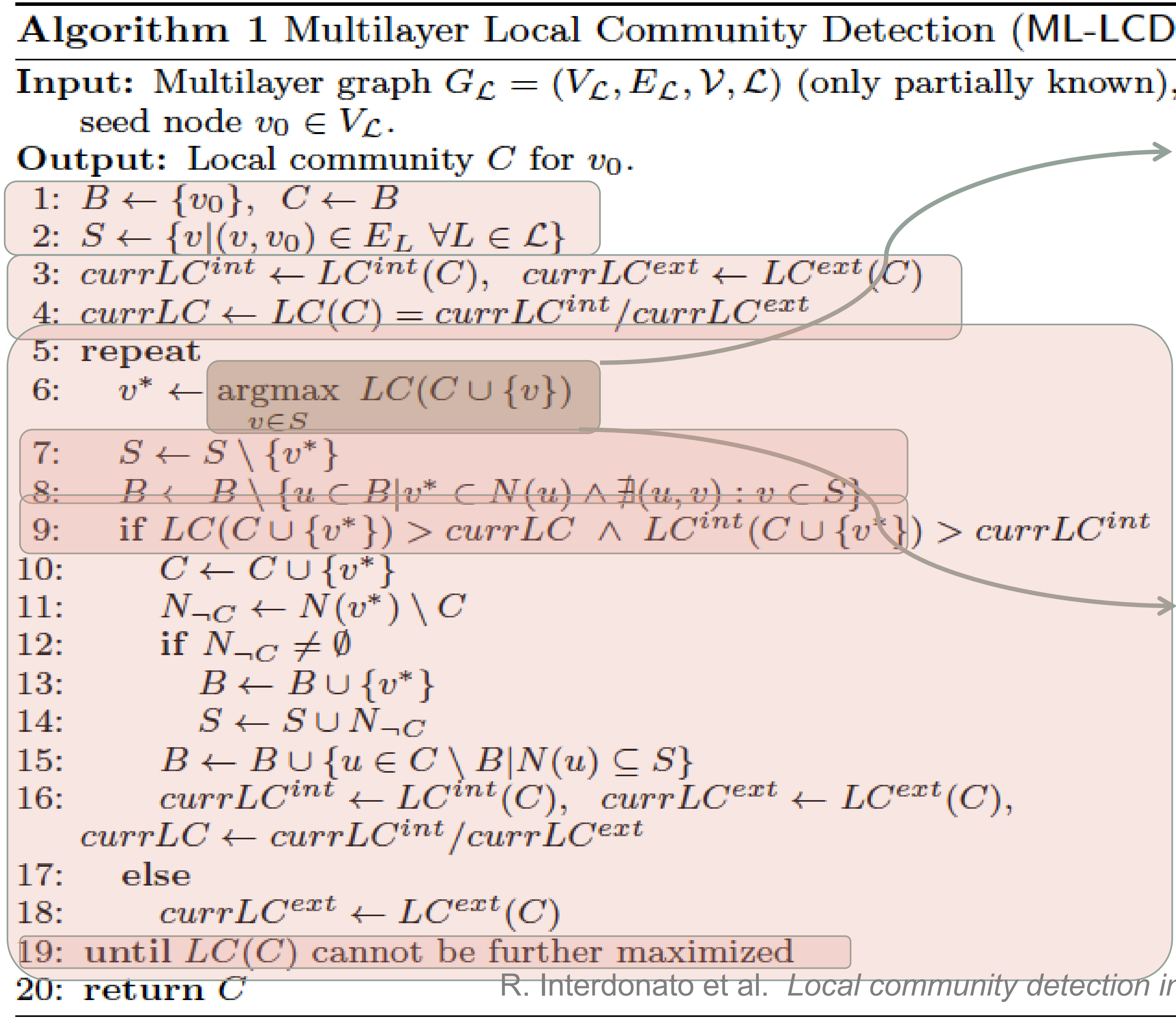
Cross-layer similarity-based local community functions



C = a,b,c,d
S = f,g,e
B = b,c,d

cl-sim:
sim(b,f)
sim(b,g)
sim(c,g)
sim(c,e)
sim(d,e)
sim(b,e)
sim(c,f)
sim(d,f)
sim(d,g)

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)



Algorithm 1 Multilayer Local Community Detection (ML-LCD) scheme

Input: Multilayer graph $G_{\mathcal{L}} = (V_{\mathcal{L}}, E_{\mathcal{L}}, \mathcal{V}, \mathcal{L})$ (only partially known), seed node $v_0 \in V_{\mathcal{L}}$.

Output: Local community C for v_0.

```
$B \leftarrow \{v_0\}, \; C \leftarrow B$
$S \leftarrow \{v | (v, v_0) \in E_L \; \forall L \in \mathcal{L}\}$
$currLC^{int} \leftarrow LC^{int}(C), \; currLC^{ext} \leftarrow LC^{ext}(C)$
$currLC \leftarrow LC(C) = currLC^{int} / currLC^{ext}$
repeat
    $v^* \leftarrow \operatorname{argmax}_{v \in S} LC(C \cup \{v\})$
    $S \leftarrow S \setminus \{v^*\}$
    $B \leftarrow B \setminus \{u \in B | v^* \in N(u) \wedge \nexists (u, v) : v \in S\}$
    if $LC(C \cup \{v^*\}) > currLC \;\wedge\; LC^{int}(C \cup \{v^*\}) > currLC^{int}$
        $C \leftarrow C \cup \{v^*\}$
        $N_{\neg C} \leftarrow N(v^*) \setminus C$
        if $N_{\neg C} \neq \emptyset$
            $B \leftarrow B \cup \{v^*\}$
            $S \leftarrow S \cup N_{\neg C}$
        $B \leftarrow B \cup \{u \in C \setminus B | N(u) \subseteq S\}$
        $currLC^{int} \leftarrow LC^{int}(C), \; currLC^{ext} \leftarrow LC^{ext}(C),$
    $currLC \leftarrow currLC^{int} / currLC^{ext}$
    else
        $currLC^{ext} \leftarrow LC^{ext}(C)$
until $LC(C)$ cannot be further maximized
return $C$
```

ML-LCD-*lw* and ML-LCD-*wlsim*:

$$LC_v^{int} = |C| LC^{int} + \sum_{u \in C} \sum_{\substack{L_i \in \mathcal{L} \\ \wedge (u,v) \in E_i^C}} \Gamma$$

$$LC_v^{ext} = |B| LC^{ext} + \sum_{u \in S_v} \sum_{\substack{L_i \in \mathcal{L} \\ \wedge ((u,L_i),(v,L_i)) \in E_{\mathcal{L}}}} \Gamma - \sum_{u \in B_v} \sum_{\substack{L_i \in \mathcal{L} \\ \wedge (u,v) \in E_i^B}} \Gamma$$

ML-LCD-*clsim*:

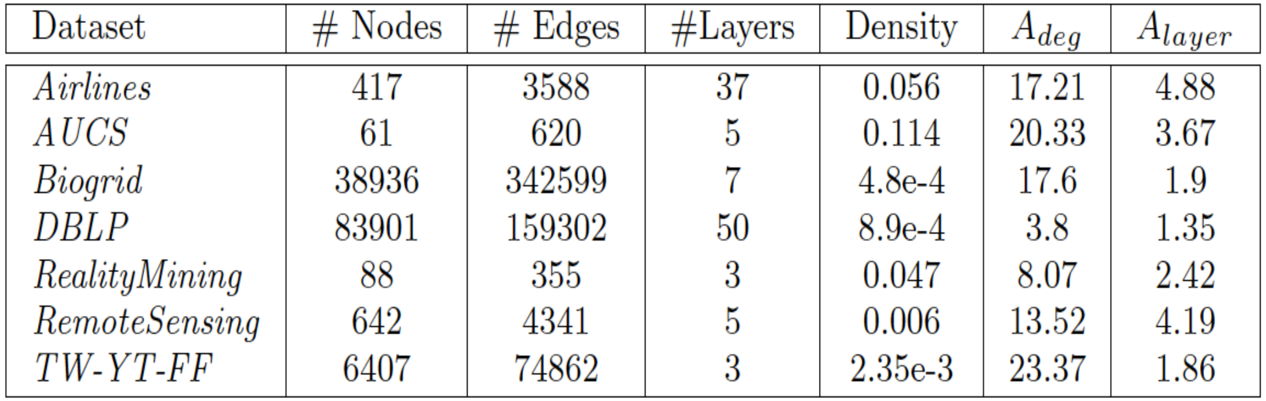
$$LC_v^{int} = |C| LC^{int} + \sum_{u \in C} \sum_{\substack{L_i, L_j \in \mathcal{L} \\ \wedge ((u,L_i),(v,L_j)) \in E_{\mathcal{L}}}} sim_{i,j}(u,v)$$

$$LC_v^{ext} = |B| LC^{ext} + \sum_{u \in S_v} \sum_{s \in S} \sum_{\substack{L_i, L_j \in \mathcal{L} \\ \wedge ((u,i),(s,j)) \in E_{\mathcal{L}}}} sim_{i,j}(u,s) + $$

$$- \sum_{u \in B_v} \sum_{s \in S} \sum_{\substack{L_i, L_j \in \mathcal{L} \\ \wedge ((u,i),(s,j)) \in E_{\mathcal{L}}}} sim_{i,j}(u,s)$$

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Evaluation goals



Dataset	# Nodes	# Edges	#Layers	Density	A_{deg}	A_{layer}
Airlines	417	3588	37	0.056	17.21	4.88
AUCS	61	620	5	0.114	20.33	3.67
Biogrid	38936	342599	7	4.8e-4	17.6	1.9
DBLP	83901	159302	50	8.9e-4	3.8	1.35
RealityMining	88	355	3	0.047	8.07	2.42
RemoteSensing	642	4341	5	0.006	13.52	4.19
TW-YT-FF	6407	74862	3	2.35e-3	23.37	1.86

- Evaluation of ML-LCD methods
  - Size of extracted LCs
  - Structural characteristics of LCs
  - Similarity between LCs
  - Distribution of layers involved in LCs
  - LC distribution over number of edges
  - Overlap of LCs
  - Efficiency analysis

- Comparison with ***single-layer, local* community detection**:
  - **LCD** (Chen et al. 2009), **Lemon** (Li et al. 2015)
- Comparison with ***multi-layer, global* community detection**:
  - **PMM** (Tang et al. 2009), **GL** (Mucha et al. 2010), **LART** (Kuncheva and Montana 2015)

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Summary of evaluation of ML-LCD methods

- ML-LCD-*lw*, then ML-LCD-*clsim*, produce larger communities
  - Ordering by “xenophobic” level:

    ML-LCD-*lw* < ML-LCD-*clsim* < ML-LCD-*wlsim*
- All methods are able to produce small-world communities
- ML-LCD-*wlsim* and ML-LCD-*clsim* behave fairly similarly (Jaccard sim of LCs)
  - *cos* similarity and *triad*-based similarity more inclusive behavior than *jac* similarity
- LCs of ML-LCD-*lw* and ML-LCD-*clsim* cover all or most of the layers
- Relatively low overlap (at node level) among LCs produced by every method
- ML-LCD-*wlsim* is the most efficient method

ML-LCD-lw	$\mathcal{O}(\|C\|^2 \times d^2 \times \ell)$
ML-LCD-$wlsim$	$\mathcal{O}(\|C\|^2 \times d^3 \log d \times \ell)$
ML-LCD-$clsim$	$\mathcal{O}(\|C\|^3 \times d^3 \log d \times \ell^2)$

R. Interdonato et al. *Local community detection in multilayer networks*. Data Min. Knowl. Discov. (2017)

Neighborhood as Influence Trigger Set

Topology-driven diversity-based (targeted) influence maximization

Main references for this part

IEEE TRANSACTIONS ON
KNOWLEDGE AND
DATA ENGINEERING

A publication of the IEEE Computer Society

SPECIAL SECTION on RULE REPRESENTATION

REGULAR PAPERS

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli:
Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks.
IEEE Trans. Knowl. Data Eng.. 30(!2): 2421-2434 (2018)

Social networks and spread of influence

- Social influence
  - A causal process: individual *u* exerts a ”force” on individual *v* to introduce a **change of the behavior** of *v*
  - Assumes existence of (online) connections and consequent interactions between individuals in a graph network
- “Word-of-mouth” effect
  - If one convinces a social contact to adopt an idea/opinion/product, then the persuaded individual will endorse the idea/opinion/product among her friends

Influence Maximization at a glance –

General setting

- Given
  - a limited budget *k* for initial advertising
  - Estimates of influence between individuals
- Problem
  - Select a set *S* of *k* individuals, or seeds, s.t., by “activating” them, the **expected spread of influence** (starting from *S*) is maximized
- Operational goal
  - Choose a diffusion model, Trigger a large cascade of influence, i.e., further adoptions of a product/info/idea
- Applications
  - Viral marketing, epidemics, recommendation, trust propagation, etc.

D. Kempe, J. M. Kleinberg, and E. Tardos. *Maximizing the spread of influence through a social network*. In Proc. ACM KDD 2003

Targeted Influence Maximization

- Focus on a selection of individuals (rather than the entire social network) through which the spread of influence
  - e.g., an organization often wants to narrow the advertisement of its products to users having certain needs or preferences, as opposed to targeting the whole crowd
  - e.g., in an OSN scenario, some events or memes would be of interest only to users with certain tastes or social profiles

Leveraging diversity for enhanced IM

- **Diversity** as a key-enabling dimension in data analysis
  - to enhance productivity,
  - to develop wiser crowdsourcing processes,
  - to improve user satisfaction in content recommendation based on novelty and serendipity,
  - to avoid “information bubble” effects, and
  - to handle legal and ethical implications in information processing

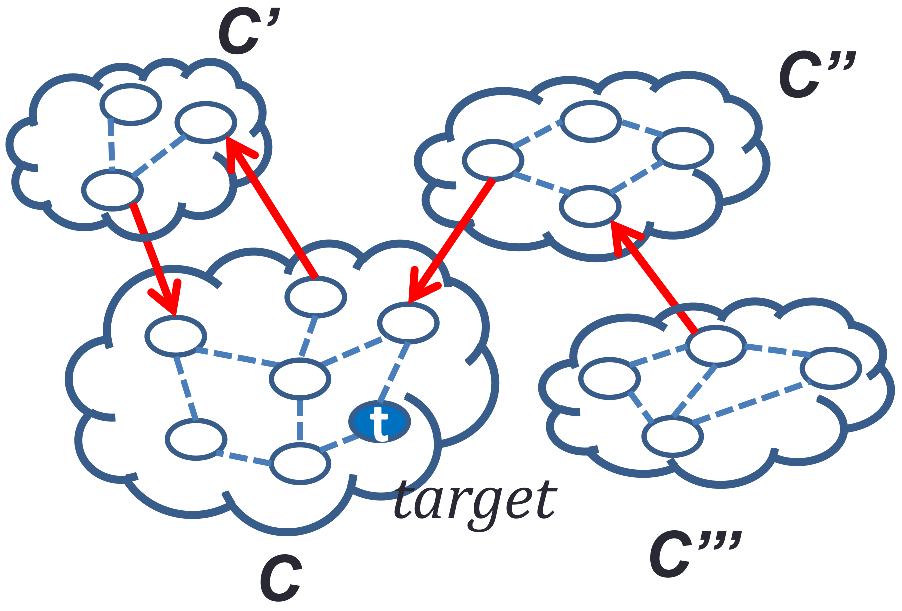
Leveraging diversity for enhanced IM

- The success of an IM task might depend not only on the size of the seed set,
- but also on the diversity that is reflected *within*, or *in relation to*, the seed set
- Diverse individuals tend to connect to others with many different characteristics
  - Personal profile, e.g., topical preferences
  - Community role(s)
  - Strategic location

How does this relates to the ability of targeting individuals?

Community-based targeted IM

- Different constraints to select community-aware portions of the diffusion graph:
  - The community graph (***C***)
  - The **weakly-knit expanded** community
    - {***C***, ***C'***, ***C''***}
  - The **tightly-knit expanded** community
    - {***C***, ***C'***}
  - The **D-recursively expanded** community
    - {***C***, ***C'***, ***C''***}
    - {***C***, ***C'***}



R. Interdonato, C. Pulice, A. Tagarelli. *Community-based delurking in social networks.* In Proc. IEEE/ACM ASONAM 2016

Community-based targeted IM

- **RQ**: Are the target users activated from seed users that belong to the **same** community, or to **external** communities?

- The best seeds are more likely to be identified among members of communities external to that of the targets
  - Such external communities are actually adjacent (or distant few hops) to the community containing the targets, linked through *bridges*
- By expanding the diffusion context (i.e., the community subgraph), it is more likely to engage targets having high activation probability

R. Interdonato, C. Pulice, A. Tagarelli. *Community-based delurking in social networks.* In Proc. IEEE/ACM ASONAM 2016

Diversity: “diverse” notions

- Common (more intuitive) notions would rely on side-information or a-priori knowledge on user attributes
- Little research on that
  - Tang et al. 2014
    - consider side-information-based diversity assuming numerical representation of node attributes, given a predetermined set of types
    - a linear combination of the expected spread function and a numerical attribute-based diversity has to be maximized by means on heuristic search strategies, defined upon classic centrality measures

F. Tang, Q. Liu, H. Zhu, E. Chen, and F. Zhu. *Diversified social influence maximization*. In Proc. IEEE/ACM ASONAM 2014

Diversity: “diverse” notions

- Different perspective adopted in Caliò et al. 2018:

  a user’s diversity in a social graph can be determined based on topological properties related to her/his neighborhood

- It just requires topology information only!

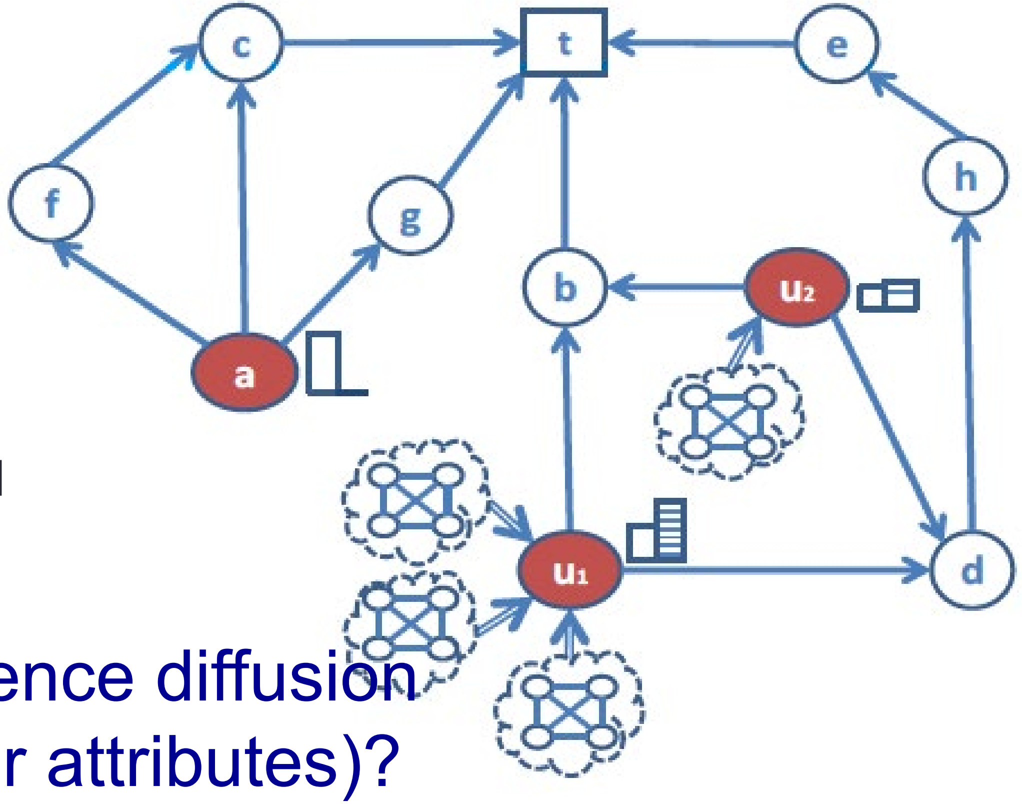
- Finds justification in Social Embeddedness and Boundary Spanning theories
  - *OSN users may naturally get knowledge from some of their social contacts and then spread the acquired capital to other contacts, spanning it from one or more components of the social graph to others*

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Main hypothesis:

- if we learn seeds that are not only capable of influencing but also are linked to more diverse (groups of) users, then we would expect that the influence triggers will be diversified as well
  - i.e., target users will get higher chance of being activated



RQ1: How to determine diversity in an influence diffusion graph (having no a-priori knowledge on user attributes)?

RQ2: How to learn the seeds by also considering diversity w.r.t. a target set?

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

- Two alternative ways of modeling topology-driven diversity
  - Depend on the approach adopted to exploit structural information from the diffusion subgraph specific to a given target node
  - **Local diversity**:
    - captures the likelihood of reaching it from nodes outside the currently unfolded target-specific diffusion subgraph
    - computed at each step of the expansion of a target-specific diffusion subgraph
  - **Global diversity**:
    - determines the diversity of nodes that lay on the boundary of the subgraph, i.e., nodes that can receive influence links from nodes external to the subgraph

      → boundary-spanning effect of external sources of influence
    - exploits the structure of the fully unfolded target-specific diffusion subgraph

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Basic definitions

- Social graph: $\mathcal{G}_0 = \langle \mathcal{V}, \mathcal{E} \rangle$
- Diffusion graph: $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, b, \ell \rangle$
  - With b edge-weighting function and ℓ node-weighting function
- Target-specific diffusion subgraph: $G_t^{(\tau)} = \langle V_t, E_t \rangle \subseteq \mathcal{G}_0$
  - a DAG rooted in the target node *t* corresponding to the portion of involved in the diffusion towards *t* at time τ
- Boundary set $B_t^{(\tau)}$ nodes having at least one incoming connection from nodes in $\mathcal{G}$ outside $G_t^{(\tau)}$
- Expansion of $G_t^{(\tau)}$: the graph $G_t^{(\tau+1)}$ resulting from the reverse unfolding of $G_t^{(\tau)}$ to contain nodes that can reach those in $B_t^{(\tau)}$
- In-neighbors of $v \in B_t$ that are not linked to v in $G_t^{(\tau)}$: $N_{\neg E_t}^{in}(v)$

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Local diversity

- Given the currently unfolded G_t and node $v \in B_t$ with $N^{in}_{\neg E_t}(v) \neq \emptyset$.
- To determine the local diversity of any node u in $N^{in}(v)$ s.t.

- *The diversity of u should be proportional to the likelihood of reaching it from nodes outside G_t .e., proportional to the number of u in-neighbors in $\mathcal{G}$ not already in G_t .*
- *The diversity of u should be proportional to the increment contributed by that node to the number of incoming links not already included in G_t*

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Local diversity

- Local diversity of u:

$$div_t(u) = \frac{\delta_t^{+u}}{\delta_t} = \frac{|B_t|}{1+|B_t|}\left(1 + \frac{|N^{in}_{\neg E_t}(u)|}{\sum_{v \in B_t} |N^{in}_{\neg E_t}(v)|}\right)$$

- i.e., the *boundary diversity conditional on inclusion* of u in G_t:

$$\delta_t^{+u} = \frac{|B_t|\delta_t + |N^{in}_{\neg E_t}(u)|}{|B_t|+1}$$

- divided by the actual *boundary diversity*:

$$\delta_t = \frac{1}{|B_t|} \sum_{v \in B_t} |N^{in}_{\neg E_t}(v)|$$

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Global diversity

- G_t is here regarded as the fully expanded diffusion graph for target *t*

- *The boundary spanning should be regarded as exogenous to the diffusion process for a specific target, i.e., associated to external sources of influence coming from the rest of the social graph*

- Boundary diversity of $v \in B_t$: the contribution of v to the boundary diversity δ_t

$$div_t^B(v) = \frac{|N_{\neg E_t}^{in}(v)|}{|B_t|}$$

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity in (targeted) IM

Global diversity

- Consider also the outward connectivity of boundary nodes:

$$|N_{E_t}^{out}(v)|/|B_t|$$

- To maximize diversity of nodes that propagate towards a given target, *the diversity of a boundary node should be*
  - *Proportional to its boundary diversity*
  - *Proportional to its outward internal span*

- Global diversity of v:

$$div_t(v) = div_t^B(v) \times \mathrm{f}\left(\frac{|N_{E_t}^{out}(v)|}{|B_t|}\right)$$

  - with **f** smoothing function to assign the outward internal span a weight at most equal to the boundary diversity, i.e., $\mathrm{f} = \log(1 + |N_{E_t}^{out}(v)|/|B_t|)$

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)

Topology-driven diversity-sensitive targeted IM

Complexity and algorithms

- DTIM maintains the complexity of IM problems, but
- Good news:
  - Both capital function and local/global diversity function are proven to be nondecreasing **monotone** and **submodular**
  - And so is the DTIM objective function (Linear Threshold model is used)
- Therefore, a greedy algorithm can be developed with (1 – 1/*e*) approximation

- Original DTIM algorithms follow a greedy approach that exploits the search for shortest paths in the diffusion graph, in a backward fashion from the selected target set
- Also available: state-of-the-art Reverse-Influence-Sampling-based formulation of DTIM

A. Caliò, R. Interdonato, C. Pulice, A. Tagarelli. *Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks*. IEEE TKDE (2018)