2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

NI-MLA: Node Importance based Multi-level Label Assignment strategy for community detection in sparse social graphs

1st Elyazid Akachar and 2nd Yahya Bougteb
National Higher School of Arts and Crafts (ENSAM)
Moulay Ismail University, Meknes, Morocco
elyazid.akachar@gmail.com
Y.bougteb@edu.umi.ac.ma

3rd Meriem Adraoui
Center of Urban Systems (CUS)
Mohammed VI Polytechnic University (UM6P)
Benguerir, Morocco
meriem.adraoui@um6p.ma

4th Brahim Ouhbi
National Higher School of Arts and Crafts (ENSAM)
Moulay Ismail University, Meknes, Morocco
b.ouhbi@umi.ac.ma

5th Bouchra Frikh
LIASSE Lab,National School of Applied Sciences (ENSA)
Sidi Mohamed Ben Abdellah University, Fez, Morocco
bouchra.frikh@usmba.ac.ma

***Abstract*—This research paper addresses the challenge of detecting communities in sparse social graphs and presents a novel approach that leverages node importance and label propagation. The proposed method consists of three phases: initialization, label assignment, and filtering. In the initialization phase, we carefully identify and designate key nodes using their local information and associate them with different labels. Subsequently, in the label assignment phase, the assigned labels are propagated to neighboring nodes, which are organized in a multilevel manner, taking into account their relevance and significance. Through the filtering phase, we effectively eliminate irrelevant labels, enhancing the accuracy of community assignments and resulting in an optimized community structure. To assess the effectiveness of our approach, we conducted experiments on both real-world networks and synthetic networks. A comparative analysis was performed against several established community detection techniques from existing literature. The results clearly demonstrate that our proposed algorithm surpasses existing methods in terms of accuracy and efficiency.**

***Index Terms*—Community detection, node importance, label propagation, social networks**

I. Introduction and related work

Real-world systems across various fields consist of interacting entities, forming networks. Typically, these networks are depicted as graphs with nodes as entities and edges indicating interactions. An example is social networks, which can be analyzed using graph theory tools. Analyzing network properties is vital to understanding information within these systems.

One major problem to address while examining social networks lies in the identification of communities or clusters. Generally, communities consist of members with shared friends and common interests. [1], [2]. this paper specifically focuses on community detection in social networks. This field has seen significant advancements, resulting in diverse approaches, including graph partitioning, spectral clustering, optimization-based, and heuristic algorithms. While traditional methods like graph partitioning and spectral clustering require predefined parameters, optimization-based and heuristic methods automatically determine the number of communities based on network topology. These two categories seek to optimize objective or likelihood functions such as modularity [3]. Inspired by the mechanism of these approaches several methods have been introduced. Notable examples include Girvan-Newman (GN), Fastgreedy (FG), Louvain, Walktrap, Infomap, Label Propagation (LPA), Spinglass, and Leading Eigenvector (LEV). For more details and additional methods, refer to surveys [1], [2].

Recently, Tunali proposed SimCMR [4] algorithm that merges communities based on several refinement operations and modularity maximization to form a hierarchical structure. Ahajjam et al. [5] introduced LCDA1 and LCDA2 methods, which utilize leaders to identify cohesive communities based on eigenvector centrality. Bouyer et al. [6] presented LSMD, a community detection algorithm using local information and multi-level diffusion.

As networks grow in scale, particularly social networks,, their complexity makes community detection challenging. Traditional methods, which may rely on set community counts or maximizing modularity, can overlook small communities and influential nodes. Recognizing these influential nodes is vital

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627598

because they influence community behavior, offering valuable insights for areas like e-learning, politics, and recommendations [1].

For these purposes,researchers have explored local node features. The seed expansion (or seed-centric) approach uses nodes as seeds to extract relevant communities from large graphs. Different strategies, such as selecting interconnected nodes as seeds, have been developed to improve efficiency and accuracy in community detection [1], [6]

This research paper introduces the **NI-MLA** algorithm (Node Importance based on Multi-level Label assignment strategy for community detection), an efficient expansion model for community detection in social networks. Unlike most existing methods, our approach simultaneously identifies influential nodes (leaders) and the communities surrounding them. The model consists of three stages. In the first phase, small dense regions in the graph are selected as the leaders of communities based on the degree centrality of nodes and maximal cliques. In the second phase, a multi-level Label assignment strategy is used to expand the cluster of leaders identified previously. Finally, a filtering process is applied to eliminate irrelevant labels and small communities to obtain an optimal community structure. In each phase, the nodes are processed according to their importance. The main idea behind our proposal is outlined in Figure 1 .The primary contributions of our research can be outlined as follows:

- We introduce NI-MLA, a novel algorithm for detecting communities and their influential nodes.
- Our approach employs a multi-level label assignment strategy to construct an initial community structure, followed by a filtering phase to optimize the obtained community assignments. The propagation process involves computing the importance of labels and nodes within the graph, leading to accurate community detection.
- Experimental results on real-world and synthetic networks demonstrate the efficiency and effectiveness of NI-MLA, outperforming widely used algorithms in discovering communities in social networks.

The rest of this paper is structured as follows: Section II is the core of the paper it presents the algorithm and necessary concepts. The evaluation and experimental results of the proposed method are reported in section III. While the conclusion is given in section IV.

II. Proposed method: NI-MLA method

In this section, we will outline the various steps of our proposed approach and discuss the underlying fundamental concepts on which it is based. Each step will be described in detail, highlighting the guiding ideas and principles.

A. Background

This section presents essential contextual information derived from the fields of graph theory and linear algebra.

Problem formulation. Given a social network that contains a set of users $U = \{u_1, u_2, u_3, ..., u_n\}$ with m interactions. We seek to divide the set of members U into a set $P = \{C_1, C_2, ..., C_p\}$ of disjoint communities in which each user $u_i \in U$ belongs only to one community. More formally, $P = \{C : C \subset U, |C| > 1\}$ such that $\forall C_i, C_{j \neq i} \in P : C_i \cap C_j = \emptyset$. The main objective of this work is to find a configuration for P where the users of each community $C_i \in P$ are strongly interconnected and weakly connected to other community users $C_{j \neq i} \in P$. Furthermore, for each community $C_i \in P$ we identify its associate influential users (or leaders).

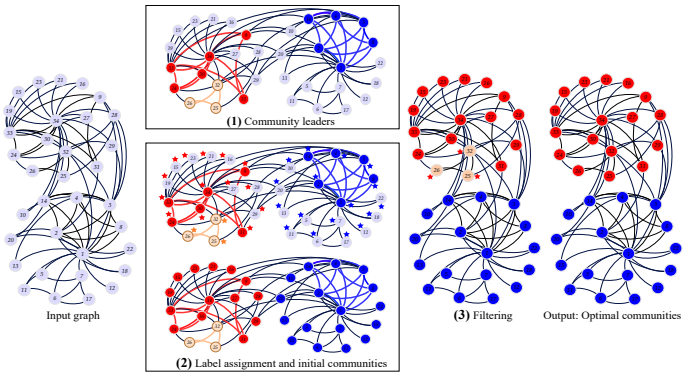

Fig. 1: The primary steps of the NI-MLA algorithm.

Definition 1 (Graph)**.** In the context of graph theory, a graph G is defined as an ordered pair $G = (V, E)$, where: V, also represented as $V(G)$, is a set of vertices within the graph G. It is important to note that V must not be an empty set ($V \neq \emptyset$). E, also referred to as $E(G)$, represents a set of edges within the graph G. Specifically, the set E is a subset of V^2, which indicates that each edge is defined as an ordered pair of vertices from the set V. $|V|$ and $|E|$ respectively give the number of nodes and edges in the graph G.

Definition 2 (Degree of vertex)**.** For an undirected graph $G = (V, E)$, the degree of a vertex $u \in V$,denoted as $d_G(u)$, is the number of edges incident to u.

Definition 3 (Clique)**.** In graph theory, a clique-denoted as Cl- is a subset of vertices within a graph where every vertex is directly connected to every other vertex in the subset. In other words, a clique is a complete sub-graph where all vertices are mutually adjacent (i.e., all vertices in a clique are fully connected to each other). More formally:

$$Cl \subset V, \forall v_i, v_j \in Cl \Rightarrow e_{v_i, v_j} \in E$$

Definition 4 (k-Clique)**.** In graph theory, a k-clique-denoted as $k - Cl$- is a clique of size equals to k (i.e., clique contains k vertices $|Cl| = k$).

The size of a clique is determined by the number of vertices it contains. For example, a clique of size 2 is called an edge, a clique of size 3 is a triangle, and so on.

Definition 5. In graph $G = (V, E)$, a maximal clique is a particular clique that cannot be expanded by adding further adjacent nodes without modifying its distribution.

Definition 6 (maximum clique)**.** A maximum clique-denoted by Cl_{max}- is the largest maximal clique in graph $G = (V, E)$ (maximum cliques are clearly maximal).

B. Methodology

After defining the necessary notions, in this subsection, we explain the main idea behind our proposal. Overall the NI-MLA method consists of three steps: Community' leaders identification, multi-level label assignment, and filtering. In what follows we describe each phase.

1) Phase 1: Community' leaders identification. In this part, we will describe the initial phase of our algorithm, primarily focusing on two key concepts: node degree and maximum cliques.

Question 1: Who holds the greatest level of influence in a social network?

In social networks, a user's influence is determined by their interactions. Therefore, to measure a user's influence, we can calculate the degree of the node representing that user, as defined by vertex degree (Definition 2). A higher degree indicates more connections and greater influence within the graph. selecting the most influential nodes in a graph can result in choosing only a single node as the most influential, which may not be useful for community detection, especially in sparse graphs. To address this limitation and select a more diverse set of candidate influential users in social networks, this work proposes an alternative approach. Instead of choosing a small subset, all nodes that have a degree greater than the average degree of the nodes are selected as candidate influential nodes. By considering nodes with above-average degrees, a broader range of potentially influential users is considered, allowing for a more comprehensive exploration of influential individuals within the social network. This approach aims to capture a larger pool of candidates for further analysis and community detection purposes. Consequently, the set of most influential nodes in a graph $G = (V, E)$ denoted by **inf** is:

$$inf(G) = \{v_i \in V | d_G(v_i) > d_{avg}(G)\} \tag{1}$$

where $d_{avg}(G)$ denotes the average degree of graph's nodes which calculated by:$d_{avg}(G) = \frac{1}{n}\sum_{i=1}^{n} d_G(v_i)$

The literature provides various methods to measure individual influence in social networks. This work chooses a specific measure balancing speed and result quality. Though there might be some quality trade-offs, the emphasis is on swift efficient real-world network processing. The degree centrality measure is favored here because of its quick linear time complexity in relation to the network's node count $\theta(n)$.

Question 2: Is having a vast number of relationships in social networks sufficient for community formation?

Influential nodes in social networks can effectively spread information due to their many relationships, making them popular. However, sheer popularity doesn't ensure broad information dissemination. While leaders can't reach large audiences alone, they rely on closely connected "assistants" to help spread information. These assistants, interconnected and directly linked to the leader, are crucial in information dissemination, as seen in political networks where a party chairperson leans on assistant members to share the party's ideologies and projects. For these purposes, we use the concept of maximum clique presented in Definition 6 to select the leaders (i.e, influential node and its assistants) that participate in the formation of community. More precisely, Let v_i the most i^{th} influential vertex in the graph, and $Cl_{max}(v_i)$ the set of q maximum cliques that contains v_i. In the proposed method, the community' leaders are formulated in the following definition.

Definition 7 (community' leaders)**.** The leaders of community C_i of partition P-denoted as $L(C_i)$- is the union of maximum cliques constructed around the most i^{th} influential node. Formally:

$$L(C_i) = \bigcup_{j=1}^{q} Cl_{max}(v_i) \tag{2}$$

To sum up, the different steps involved in the community' leaders identification are ordered as follows:

1) Obtain the list of cliques "$List\_cliques$" in the graph G that have a size greater than 2.
2) Arrange the set of candidate influential nodes $inf(G)$, obtained using Equation 1, in descending order.
3) For each element v_i in the set of candidate influential nodes $inf(G)$:
   a) Extract the set of maximum cliques formed around v_i, denoted as "$Cl_{max}(v_i)$".
   b) Consider the union of all these maximum cliques, denoted as $L(v_i)$, as the leaders of community C_i.
   c) Remove $Cl_{max}(v_i)$ from the list of cliques $List\_cliques$.
   d) If no cliques are formed around v_i, move to the next element in $inf(G)$.
4) Repeat steps **3)** until all candidate influential nodes are processed.

2) Phase 2: Label assignment. In the second phase, we turn our attention to the non-leader nodes. We use a multi-level label assignment approach to expand their local communities. Initially, nodes within each leader's group share the same label, which is then passed to non-leader nodes. This prompts growth in their communities, but an intriguing question arises:

Question 3: Who are the targeted members of the leaders?

In real-world scenarios, leaders often prioritize popular non-leader members in the network. This strategic approach allows them to rapidly spread their ideas and messages through the network by harnessing the influence of these popular individuals. In the second phase, non-leader nodes are managed based on their influence within the graph. These nodes are sorted by degree in descending order and divided into different levels, each representing a specific degree range (e.g., Level 1 for degree d_1, Level 2 for degree d_2, and so on). The nodes in each level, should be select the group of leaders they want to join, guided by a function measuring the importance of the label to a given vertex, defined as follows:

$$f(l_i, v) = \frac{|N_v(l_i)|}{d_G(v)} + \left(1 - \prod_{i \in N_v(l_i)} \frac{1}{e^{d_G(i)}}\right) \quad (3)$$

Where $N_v(l_i)$ represents the set of neighbors of a node v that are labeled with label l_i ($|N_v(l_i)|$ is its size), while $d_G(v)$ refers to the degree of node v in the graph G. This function takes into account both the importance of the label and its location. Frequent labels are deemed more important, and label propagation probability varies with location, gauged by the node's degree. In this context, the non-leaders nodes select the label that maximize the function f in Equation 3.

Overall, the different steps of label assignment phase are ordered as follows: **(1)** The nodes within each leaders group, which were obtained during the first phase, are assigned with the same label. Conversely, the non-leader nodes are left without any label. **(2)** The non-leader nodes are categorized into different levels based on their degree, with each level consisting of nodes of the same degree. These nodes are then processed based on their importance, starting from the nodes in the first level and proceeding to the nodes in subsequent levels until all nodes are processed. **(3)** Each non-leader node v selects its suitable label based on Equation 3, where the label with the highest value of f is assigned to node v.

3) Phase 3: Filtering. The main objective of this phase is to identify and eliminate certain communities that are formed around weak leaders. Weak leaders are characterized by their inability to expand their community and effectively propagate their ideas. To accomplish this, we employ the concept of propagation frequency, which quantifies the frequency at which each label is propagated. Labels with low propagation frequency are considered weak, and consequently, the corresponding leaders' groups are identified as weak leaders. The propagation frequency can be calculated as follows:

$$P_f = \frac{N(l_i)}{P} \quad (4)$$

Here, $N(l_i)$ represents the number of times label l_i is propagated, and P denotes the total number of possible propagations, which is equal to the number of non-leader nodes. In our proposed approach, any label that has a propagation count lower than the propagation frequency is considered weak. Once the weak leader groups are identified, they are treated as non-leader nodes, and their new alternative labels are selected using Equation 3.

III. Experimental results

After outlining the various steps of our proposed method, this section focuses on evaluating its performance in both real-world and synthetic networks. To assess the quality of the obtained communities, we utilize three commonly used measures: NMI, ARI , and modularity [3]. For baseline methods, we compare our proposal with several models from the literature including Louvain, Fastgreedy, Griven-Newman (GN), Walktrap, Spinglass, Leading eigenvector (LEV), LCDA1, LCDA2 and SimCMR. For more details about these methods refer to [2], [4], [5]

TABLE II: Modularity values and number of communities detected by each algorithm on real-world networks.

	Karate		Dolphin		Polbooks		Football		School		Thiers	
	N_c	Q	N_c	Q	N_c	Q	N_c	Q	N_c	Q	N_c	Q
NI-MLA	2	O.37	3	0.49	2	0.50	7	0.55	10	0.37	9	0.57
Louvain	4	0.41	5	0.52	4	0.53	10	0.60	6	0.40	6	0.58
Spinglass	4	0.41	5	0.52	6	0.52	11	0.60	5	0.39	7	0.58
Walktrap	5	0.35	4	0.48	4	0.50	10	0.60	6	0.37	7	0.58
LPA	3	0.32	6	0.49	4	0.48	11	0.58	1	0	8	0.57
GN	5	0.40	5	0.51	5	0.51	10	0.59	7	0.33	7	0.57
FG	3	0.38	4	0.45	4	0.50	6	0.54	3	0.34	5	0.54
LEV	4	0.39	5	0.49	4	0.46	8	0.49	5	0.36	4	0.47
SimCMR	3	0.39	4	0.52	5	0.51	9	0.60	7	0.39	4	0.58
LCDA1	2	0.22	4	0.21	3	0.48	8	0.52	8	0.25	11	0.51
LCDA2	5	0.08	5	0.28	3	0.48	9	0.50	8	0.25	10	0.47

A. *Experiments on real networks: Results and discussion*

For real world networks, we adopt six graphs of different sizes described in Tables I. The results obtained in real networks are reported in Figure 2 and Table II.

In the karate club network, NI-MLA successfully identifies two communities with perfect matches to the ground truth (NMI and ARI both at 100%), outperforming other methods. As to the dolphin networks, it detects three communities, closely resembling the ground truth with NMI and ARI values greater than those of its counterparts (NMI=0.88,ARI= 0.79). For Polbooks networks, our proposal identifies three communities with high NMI and ARI scores compared to other methods. Regarding the football network, NI-MLA finds seven communities, performing well in NMI and ARI but slightly lagging behind due to the graph's density. In the School and Thiers graphs, it matches the real number of communities with NMI and ARI values consistently equal to or greater than 0.90, surpassing other methods.

Table II presents modularity results, where NI-MLA consistently achieves competitive values exceeding 0.40, despite not being designed for modularity maximization. NI-MLA typically ranks among the top three or four algorithms, alongside Louvain, Spingalss, and SimCMR. To sum up, the NI-MLA algorithm demonstrates a strong ability to detect significant communities within real-world networks,

B. *Experiments on synthetic networks: Results and discussion*

Synthetic networks are commonly used as a benchmark to assess community detection methods' accuracy, with the LFR (Lancichinetti–Fortunato–Radicchi) model being a widely accepted choice [13]. We employed LFR-generated benchmark graphs to evaluate methods using the NMI measure. Graphs generated under the LFR model have several adjustable parameters, with the mixing parameter $\mu \in]0, 1[$ being the most important one. As the value of μ increases, the ambiguity of the community structure also increases. Details on other parameters and their settings are in Table III. Notably, we excluded GN, LCDA1, and LCDA2 methods from these experiments due to their time-consuming nature.

A uniform remark can be extracted from the figures reporting the experiments on synthetic networks (Figures 3 to

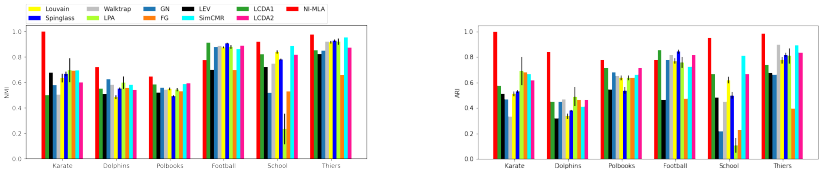

Fig. 2: NMI and ARI values obtained by the examined algorithms on real-world graphs.

TABLE I: Properties of real networks

Networks	#V	#E	#C	Ref
Karate	34	78	2	[7]
Dolphin	62	159	2	[8]
Polbooks	105	441	3	[9]
Football	115	613	12	[10]
School	238	5539	10	[11]
Thiers	327	5813	9	[12]

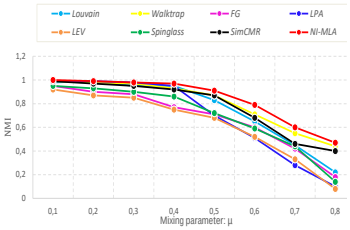

Fig. 3: LFR_1 graphs with 1000 nodes

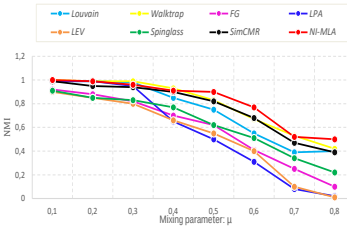

Fig. 4: LFR_2 graphf with 5000 nodes

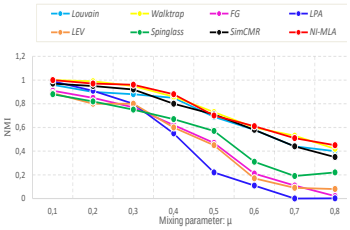

Fig. 5: LFR_3 graphs with 8000 nodes

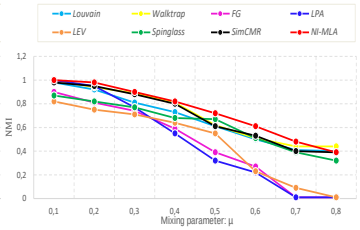

Fig. 6: LFR_4 graphs with 10000 nodes

TABLE III: The synthetic graphs and their adjustable parameters used in this study.

Network	N	K^{min}	K^{max}	C^{min}	C^{max}	τ_1	τ_2	μ
LFR_1	1000	20	50	10	50	2	1	$0.1-0.8$
LFR_2	5000	20	50	10	50	2	1	$0.1-0.8$
LFR_3	8000	20	50	20	100	2	1	$0.1-0.8$
LFR_4	10000	20	50	20	100	2	1	$0.1-0.8$

6). As the value of μ increases from 0.1 to 0.8, discovering accurate community structures becomes relatively more difficult (NMI converge to 0). This poses a greater challenge for community detection algorithms, highlighting the differences in their performance and ability to handle complex and ambiguous network structures. For a LFR_1 graphs (Figure 3), NI-MLA performs well even when $\mu = \{0.7, 0.8\}$, while other algorithms struggles. In LFR_2 graphs (Figure 4), NI-MLA maintains high NMI values across all μ values and outperforms other methods. As to the LFR_3 graphs (Figure 5), NI-MLA and Walktrap give best results. Regarding LFR_4 (Figure 6), our method outperforms others across all ambiguity levels.

To conclude, in synthetic network experiments, NI-MLA excels, followed by Walktrap, SimCMR, and Louvain, while FG, LPA, LEV, and Spinglass struggle with higher ambiguity.

IV. Conclusion and Future Works

In conclusion, this paper presents NI-MLA, a novel method for social network community detection. NI-MLA effectively identifies communities by combining node importance and a multi-level label assignment approach. It operates in three phases: initialization identifies dense regions, label assignment constructs the initial community structure, and filtering optimizes it by removing weak communities. Experimental results demonstrate NI-MLA's superior accuracy and efficiency compared to existing methods, advancing community detection in social networks. This work also suggests future directions, such as extending NI-MLA to detect overlapping communities, offering insights into complex network relationships and fostering new opportunities for network analysis.

References

[1] M. Azaouzi, D. Rhouma, and L. Ben Romdhane, "Community detection in large-scale social networks: state-of-the-art and future directions," *Social Network Analysis and Mining*, vol. 9, no. 1, Dec. 2019, publisher Copyright: © 2019, Springer-Verlag GmbH Austria, part of Springer Nature.

[2] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75 – 174, 2010.

[3] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Comput. Surv.*, vol. 50, no. 4, aug 2017.

[4] V. Tunali, "Large-scale network community detection using similarity-guided merge and refinement," *IEEE Access*, vol. 9, pp. 78 538–78 552, 2021.

[5] S. Ahajjam, M. El Haddad, and H. Badir, "A new scalable leader-community detection approach for community detection in social networks," *Social Networks*, vol. 54, pp. 41 – 49, 2018.

[6] A. Bouyer and H. Roghani, "Lsmd: A fast and robust local community detection starting from low degree nodes in social networks," *Future Generation Computer Systems*, vol. 113, pp. 41–57, 2020.

[7] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[8] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, Sep 2003.

[9] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, ser. LinkKDD '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 36–43.

[10] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, Feb 2004.

[11] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, "High-resolution measurements of face-to-face contact patterns in a primary school," *PLOS ONE*, vol. 6, no. 8, pp. 1–13, 08 2011.

[12] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PLOS ONE*, vol. 10, pp. 1–26, 09 2015.

[13] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, vol. 78, p. 046110, Oct 2008.