A Local Search Method for Graph Clustering Heuristics based on Partitional Distribution Learning

Diana Manjarres*, Itziar Landa-Torres* and Javier Del Ser*,†,‡
*TECNALIA RESEARCH & INNOVATION, 48160 Zamudio, Bizkaia, Spain
Email: {diana.manjarres, itziar.landa, javier.delser}@tecnalia.com
†University of the Basque Country UPV/EHU, Bilbao, Bizkaia, Spain
Email: javier.delser@ehu.eus
‡Basque Center for Applied Mathematics. 48009 Bilbao, Bizkaia, Spain

***Abstract*—The community structure of complex networks reveals hidden relationships in the organization of their constituent nodes. Indeed, many practical problems stemming from different fields of knowledge such as Biology, Sociology, Chemistry and Computer Science can be modeled as a graph. Therefore, graph analysis and community detection have become a key component for understanding the inherent relational characteristics underlying different systems and processes. In this regard, distinct unsupervised quality metrics such as conductance, coverage and modularity, have upsurged in order to evaluate the clustering arrangements based on structural and topological characteristics of the cluster space. In this regard graph clustering can be formulated as an optimization problem based on the maximization of one of such metrics, for which a number of nature-inspired heuristic solvers has been proposed in the literature. This paper elaborates on a novel local search method that allows boosting the convergence of such heuristics by estimating and sampling the cluster arrangement distribution from the set of intermediate produced solutions of the algorithm at hand. Simulation results reveal a generalized better performance compared towards other community detection algorithms in synthetic and real datasets.**

I. Introduction

In the last decades graph theory has been extensively applied for representing and analyzing a wide variety of systems and environments in distinct areas such as Biology, Technology, Sociology and Computer Science. In these disciplines graph analysis has become crucial to infer information and understand the characteristics of complex networks. One of the most relevant problems addressed in graph theory when applying through real systems is the inference of a clustering or community structure, i.e. groups of nodes with many edges within the same cluster and comparatively less connections between different clusters. Such communities or clusters can be regarded as groups of nodes that share similar features and play common roles within the overall graph. For instance, in Protein-Protein Interaction (PPI) networks, communities have shown to group proteins that have a specific functionality in the cell [1]. Another example that is on the rise nowadays is the creation of virtual groups across the Internet. Each virtual group can be seen as a community or cluster that includes people with the same interests or skills. In this context, the identification of persons with similar interests allows for recommendation systems to be more efficient and better guide customers to products achieving higher sales opportunities ([2]-[3]).

One of the main differences between data clustering and community detection in graphs is the definition of the quality metric that identifies the desired structural properties of the groups' partitions. While in data clustering communities are related to sets of nodes that are compared based on a distance or similarity metric, communities in graphs are more focused on the concept of internal versus external edge density. In this regard, several distinct unsupervised quality metrics have been proposed in the literature ([4]-[5]), such as: the coverage metric [6], the modularity metric [7] and the conductance metric [8]. Among the metrics discussed above, modularity is the most widely applied one to detect the strength of communities. It compares the number of inter-community edges with the expected number of links in a random graph having the same size and distribution of nodes as the original graph. Nevertheless, all these metrics have been surpassed by the appearance of an alternative global measure of performance, coined as Surprise [9]. This metric upsurged in an attempt for overcoming a cluster resolution limit, that is, the impossibility of detecting communities below a certain size threshold that depends on the overall size of the graph. By means of using a cumulative hypergeometric distribution it is possible to calculate the probability of the distribution of links and nodes in the communities defined for the network by a given partition. Therefore, Surprise metric measures how unlikely or "surprising' is that distribution.

By means of the maximization of the Surprise metric, authors in many research works have achieved outstanding results and more accurate community structure characterization than modularity-based methods over standard benchmarks and real-world scenarios ([9]-[11]). In this regard, meta-heuristic algorithms can be seen as an efficient tool for their application in communities detection in graphs due to their intrinsic capability to adapt and search near optimal solutions in complex optimization problems. Authors in [12] have recently proposed a novel heuristic community detection approach based on the so-called Firefly Algorithm (`FA`, [13]). Simulation results demonstrated that the proposed solution generalizes better than other schemes in different synthetic scenarios.

Following the same line of research, this paper builds upon

978-1-5090-4601-0/17/$31.00

this state of the art by presenting a novel local search method for graph clustering based on population-based distribution learning. More specifically, the proposed local search method constructs a matrix based on the co-occurrence of vertices in clusters of the input partitions. This matrix is then used as an input for the graph clustering technique (based on FA) leading to a new set of potentially more optimal partitions [14]. The proposed technique is tested over synthetic and real-world scenarios and is proven to outperform community detection algorithms in terms of generalized performance.

The rest of the paper is structured as follows: Section II presents the optimization problem of communities detection in graphs. Next, Section III and subsections therein describes the proposed algorithm and local search method for communities detection. Section IV presents the obtained simulation results and the comparison towards other community detection algorithms of the literature. Finally, Section V concludes the paper and proposes future research lines.

II. Graph Clustering Problem Formulation

In general, a graph $G = (\mathcal{V}, \mathcal{E})$ can be viewed as a set $\mathcal{V}$ of $V \doteq |\mathcal{V}|$ vertices or nodes, which are connected by a set of links or edges $\mathcal{E} = \{(u, v)|u, v \in \mathcal{V}\}$ each associated with a weight $w(u, v) \in \mathbb{R}$. $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ represents a clustering arrangement of the graph G in which nodes of the graph are partitioned in N disjoint groups such that $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ and $\cup_{i=1}^{N} \mathcal{C}_i = \mathcal{V}$. For the encoding of a clustering solution, i.e. the correspondence between nodes in $G(\mathcal{V}, \mathcal{E})$ and clusters in $\boldsymbol{\mathcal{C}}$, a V-sized vector of integers $\mathbf{X} = \{X_v\}_{v=1}^{V}$ with $X_v \in \{1, \ldots, N\}$ is defined.

In order to seek an optimal clustering arrangement $\boldsymbol{\mathcal{C}}^*$ an optimization problem for the maximization of the Surprise metric coined as $S(\mathcal{C})$ and formally stated in Equation (1) can be casted. The main strength of employing this metric instead of conventional ones like modularity is its capacity for being nearly unaffected by the well-known resolution limit. This alternative metric is given by

$$S(\boldsymbol{\mathcal{C}}) \doteq -\log \sum_{j=p}^{\min\{M,n\}} \frac{\binom{M}{j}\binom{F-M}{n-j}}{\binom{F}{n}}, \tag{1}$$

where $F \doteq V(V-1)/2$ represents the maximum number of links in the graph; $n \doteq \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} w(u, v)$ the actual number of links in the graph; $M \doteq \sum_{i=1}^{N} |\mathcal{C}_i|(|\mathcal{C}_i|-1)/2$ the maximum number of intra-community links for the arrangement; and p the actual number of inter-community links in the partition, computed as $p \doteq \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{u \in \mathcal{C}_i} \sum_{v \in \mathcal{C}_j} w(u, v)$. Therefore, the problem of finding the optimal clustering arrangement $\boldsymbol{\mathcal{C}}^*$ reduces to

$$\boldsymbol{\mathcal{C}}^* \doteq \arg\max_{\boldsymbol{\mathcal{C}} \in \boldsymbol{\mathcal{C}}^{\diamond}} S(\mathcal{C}), \tag{2}$$

where $\boldsymbol{\mathcal{C}}^{\diamond}$ denotes the set of all possible partitions for a V-sized graph. As stated in [15] the above problem is $\mathcal{NP}$-hard and can be solved by means of a meta-heuristic solver as presented in several research works [16], [17], [12]). The following section introduces the implemented nature-inspired solver in the latter reference – namely, a Firefly Algorithm (FA) suited to tackle this specific clustering problem –, followed by the details of the herein proposed novel local search method.

III. Proposed Algorithm

The Firefly Algorithm (FA) pioneered by Xin-She Yang in [13] consists of a population-based swarm solver inspired on the behavioral mobility patterns of fireflies. In the FA algorithm there are two important concepts: the variation of light intensity and the formulation of the attractiveness between fireflies in the swarm. For simplicity, the attractiveness of a firefly is determined by its brightness, whereas the intensity decreases as the distance to other fireflies increases. Therefore, the brightness is associated to the fitness function to be optimized, yielding the so-called update equation given by

$$\mathbf{X}_i^{t+1} = \mathbf{X}_i^t + \beta \exp[-\gamma r_{ij}^2](\mathbf{X}_j^t - \mathbf{X}_i^t) + \alpha_t \mathbf{e}_t, \tag{3}$$

where $\mathbf{X}_i^t$ represents the i-th firefly (solution) in the population at iteration t, r_{ij} denotes the distance between firefly i and j, $\mathbf{e}_t$ is a vector of random variables following a certain probability density function, and β, γ and α_t are parameters that balance between the explorative and exploitative search capabilities of the algorithm. A more detailed description of the steps of a naive FA is given in Algorithm 1.

Algorithm 1 Firefly Algorithm

```
f(X) corresponds to the fitness function to be maximized.
Generate an initial population of P fireflies {X_p^0}_{p=1}^P.
Set β, γ, α_t, the maximum number of iterations T and
    the distance function r_ij depending on the solution space
    spanned by the problem.
t = 0.
procedure FIREFLY ALGORITHM(f(X), α, β, r_ij, T)
  while t ≤ T do
    for i = 1 : P do
      for j = i + 1 : P do
        if f(X_j^t) > f(X_i^t) then,
          Vary attractiveness X_i^t with distance r via
          Equation (3).
          Evaluate new solution and update light
          intensity f(X_i^{t+1}).
        else
          Update X_j^t via Expression (3).
          Evaluate and update f(X_j^{t+1}).
        Update r_ij.
    Rank fireflies and find the best solution at iteration t
    given by the highest f(·) over {f(X_p^{t+1})}_{p=1}^P.
    t = t + 1.
```

In order to efficiently represent the differences with respect to the cluster arrangement rather than on the numerical representation of the utilized encoding, the distance among fireflies r_{ij} is given by their phenotype rather than by their genotype. Formally speaking the phenotype represents the observable

structure, whereas the genotype refers to the inheritable or internal information. Therefore, in graph clustering problems, the correspondence between nodes in $G(\mathcal{V},\mathcal{E})$ and clusters in $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1,\ldots,\mathcal{C}_N\}$ can be represented by a V-sized vector of integers $\mathbf{X} = \{X_v\}_{v=1}^{V}$ with $X_v \in \{1,\ldots,N\}$. In this paper, the operators for the proposed meta-heuristic technique are applied directly on the genotype of the problem avoiding the redesigning of the overall meta-heuristic solver. More specifically, the set of operators that are applied on the genotype space are the following:

- Step 9: the value of the Surprise metric achieved by the cluster arrangement $\boldsymbol{\mathcal{C}}$ must be maximized. As previously stated, it directly refers to the brightness of the firefly.
- Steps 10 and 13: update attractiveness by means of an increment of distance r and the *difference* $\mathbf{X}_j^t - \mathbf{X}_i^t$ between individuals. By this way, solutions with similar level of brightness should move less than solutions that lie far apart from each other in the genotype space. Moreover, the addition of a random perturbation $\alpha_t \mathbf{e}_t$ in Equation (3) allows slight changes in their phenotype space.

The approach in [12] incorporates a measure of genotypical similarity between cluster arrangements that will control the update procedure, i.e. the so-called Normalized Mutual Information score $\text{NMI}(\boldsymbol{\mathcal{C}}_i,\boldsymbol{\mathcal{C}}_j)$ [18] which is given by

$$\text{NMI}(\boldsymbol{\mathcal{C}}_i,\boldsymbol{\mathcal{C}}_j) \doteq \frac{I(\boldsymbol{\mathcal{C}}_i;\boldsymbol{\mathcal{C}}_j)}{\sqrt{H(\boldsymbol{\mathcal{C}}_i)\cdot H(\boldsymbol{\mathcal{C}}_j)}}, \tag{4}$$

where $I(\boldsymbol{\mathcal{C}}_i;\boldsymbol{\mathcal{C}}_j)$ represents the mutual information between the partitions, computed as

$$I(\boldsymbol{\mathcal{C}}_i;\boldsymbol{\mathcal{C}}_j) \doteq \sum_{n=1}^{\boldsymbol{\mathcal{C}}_i}\sum_{m=1}^{\boldsymbol{\mathcal{C}}_j} P(n,m)\log\left(\frac{P(n,m)}{P_i(n)P_j(m)}\right), \tag{5}$$

with $P(m,n) = |\mathcal{C}_{i,n}\cap\mathcal{C}_{j,m}|/V$ and $P_i(n) = |\mathcal{C}_{i,n}|/V$. It can be observed that the higher $\text{NMI}(\boldsymbol{\mathcal{C}}_i,\boldsymbol{\mathcal{C}}_j)$ is, the more similar the cluster arrangements represented by $\boldsymbol{\mathcal{C}}_i$ and $\boldsymbol{\mathcal{C}}_j$ will be.

Thus, the mutual association among the nodes within each cluster $\mathcal{C}_{i,n}$ of the attracting firefly $\boldsymbol{\mathcal{C}}_i$ will be transferred with probability $1-\text{NMI}(\boldsymbol{\mathcal{C}}_i,\boldsymbol{\mathcal{C}}_j)$) to the cluster arrangement of the attracted firefly $\boldsymbol{\mathcal{C}}_j$. This transfer is made on a node-wise basis, i.e. for every node $v \in \mathcal{C}_{i,n}$ and for every $n \in \{1,\ldots,N_i\}$ the algorithm decides, with probability $1-\text{NMI}(\boldsymbol{\mathcal{C}}_i,\boldsymbol{\mathcal{C}}_j)$, whether v in the attracted firefly $\boldsymbol{\mathcal{C}}_j$ should be associated with the rest of nodes in $\mathcal{C}_{i,n}$ or, instead, left as it was in its corresponding cluster in $\boldsymbol{\mathcal{C}}_j$. Note that if cluster arrangements aren similar to each other, the transfer probability decreases.

The random perturbation of the naive `FA` described in Algorithm 1 is implemented in the proposed algorithm in a probability basis referred to as Topological Search Rate (TSR). This parameter allows a greedily exploration of the solution space in order to encounter a topologically better assignment for the firefly at hand. Thus, the random perturbation performed at this step of the proposed algorithm will reassign node v within firefly i to cluster $\mathcal{C}_{i,n}$ with probability $TSR\cdot\Psi(v,\mathcal{C}_{i,n})/\sum_{k=1}^{N_i}\Psi(v,\mathcal{C}_{i,k})$, i.e. proportionally to the closeness centrality of the node with respect to every cluster. As stated in the following equation, this metric is the reciprocal of the sum of the shortest path distances from v to the nodes in $\mathcal{C}_{i,n}$, i.e.

$$\Psi(v,\mathcal{C}_{i,n}) = \frac{|\mathcal{C}_{i,n}|/1}{\sum_{u\in\mathcal{C}_{i,n}} d(u,v)}, \tag{6}$$

where $d(u,v)$ refers to the shortest-path distance between nodes u and v. Note that fireflies that have a great confidence about their cluster assigment will be represented by higher values of this metric.

Finally, the third operator of the algorithm evaluates the degree of *looseness* of every node $v \in \mathcal{V}$ with respect to the cluster to which it is assigned in $\boldsymbol{\mathcal{C}}_i$: if its intra-cluster degree is lower than that of the rest of the nodes in the cluster, the node is disconnected (with probability 0.5) from the cluster to form a new community. This operator allows for a discovery of small-sized clusters and ultimately, singletons.

These three operators are iteratively applied over a population of P fireflies which represent feasible cluster arrangements. The initial population is defined at random and the transfer of information is applied to every compared pair of fireflies, subsequently followed by the probabilistic topological search and the procedure to discover small-sized communities or singletons.

A. Proposed Local Search Method

Once a new set of fireflies is derived a novel local search procedure is applied to each solution to obtain an improved set of solutions. The elitism is applied at this step by selecting only those fireflies that after the application of the local search procedure the Surprise metric is increased. The proposed local seach procedure grounds on the iterative estimation at every iteration $t \in \{1,\ldots,T\}$ of a $V\times V$ probability matrix $\boldsymbol{\mathcal{D}}^t$, whose entry $\mathcal{D}_{i,j}^t$ indicates the number of partitions over the population of fireflies $\{\mathbf{X}_p^t\}_{p=1}^{P}$ in which nodes i and j of the network are assigned to the same cluster, i.e.

$$\mathcal{D}_{i,j}^t = \frac{\sum_{p=1}^{P} S(\boldsymbol{\mathcal{C}}_p)\sum_{i=1}^{V}\sum_{j=i+1}^{V}\mathbb{I}(\mathcal{C}_{p,i}^t = \mathcal{C}_{p,j}^t)}{\sum_{p=1}^{P} S(\boldsymbol{\mathcal{C}}_p)}, \tag{7}$$

where $\mathbb{I}(\cdot)$ is an auxiliary indicator function taking value 1 if its argument is true and 0 otherwise. Once composed this matrix serves as a basis for the generation of P new fireflies by randomly creating cluster arrangements that follow the pairwise association between nodes indicated by this estimated matrix. The generated solutions are then evaluated by means of the Surprise metric $S(\boldsymbol{\mathcal{C}})$ and merged with the ones obtained at previous iteration. As in other well-known evolutionary algorithms, such as Genetic Agorithm, Harmony Search or other population-based meta-heuristics, only the best solutions are selected to pass next generation. The procedure stops when a fixed number of iterations is completed.

In order to shed light on the overall procedure a more detailed description is presented in Algorithm 2. In essence the proposed local search method can be regarded as the

hybridization of a canonical Estimation of Distribution Algorithm (EDAS [27]) that constructs a probabilistic model over the genotype of the solution space as indicated by the swarm of produced fireflies at every iteration. As such the proposed local search method can be straightforward applied to any population-based global search heuristic, either from Evolutionary Computation or Swarm Intelligence.

Algorithm 2 Firefly Algorithm for graph clustering [12] with the proposed local search procedure (FA+LS)

$S(\boldsymbol{\mathcal{C}})$ denotes the fitness function to be maximized.
Initialize P fireflies $\{\boldsymbol{\mathcal{C}}_p^0\}_{p=1}^P$ at random.
Set T (# of iterations), TSR (Topological Search Rate).
Set $t = 0$.
procedure FIREFLY ALGORITHM WITH LOCAL SEARCH
  while $t \leq T$ **do**
    for $i = 1 : P$ **do**
      for $j = i + 1 : P$ **do**
        if $S(\boldsymbol{\mathcal{C}}_j^t) > S(\boldsymbol{\mathcal{C}}_i^t)$ **then**,
          Transfer node-to-cluster mappings from $\boldsymbol{\mathcal{C}}_j^t$. to $\boldsymbol{\mathcal{C}}_i^t$ with probability $1 - \text{NMI}(\boldsymbol{\mathcal{C}}_i, \boldsymbol{\mathcal{C}}_j)$, yielding $\boldsymbol{\mathcal{C}}_i^{t+1}$
          Apply the random perturbation operator over every node $v \in \mathcal{V}$ and $\boldsymbol{\mathcal{C}}_i^{t+1}$ (w.p. TSR).
          Discovery of singletons and small-sized communities in $\boldsymbol{\mathcal{C}}_i^{t+1}$.
          Evaluate and update $S(\boldsymbol{\mathcal{C}}_i^{t+1})$.
        else
          Transfer node-to-cluster mappings from $\boldsymbol{\mathcal{C}}_i^t$ to $\boldsymbol{\mathcal{C}}_j^t$ with probability $1 - \text{NMI}(\boldsymbol{\mathcal{C}}_i, \boldsymbol{\mathcal{C}}_j)$, yielding $\boldsymbol{\mathcal{C}}_j^{t+1}$.
          Apply the random perturbation operator over every node $v \in \mathcal{V}$ and $\boldsymbol{\mathcal{C}}_j^{t+1}$ (w.p. TSR).
          Discovery of singleton and small-sized communities in $\boldsymbol{\mathcal{C}}_j^{t+1}$.
          Evaluate and update $S(\boldsymbol{\mathcal{C}}_j^{t+1})$.
      Apply the proposed local search procedure to each firefly by means of matrix $\boldsymbol{\mathcal{D}}$ calculation.
      Only those fireflies that improve the $\boldsymbol{\mathcal{S}}$ metric with respect to no local search are selected.
      $t = t + 1$.
  The best cluster arrangement at iteration t is given by the highest $S(\cdot)$ among $\{S(\boldsymbol{\mathcal{C}}_p^T)\}_{p=1}^P$.

IV. Simulation Results and Discussion

This section presents the performance of the proposed graph clustering approach over synthetic and real-world network topologies with varying sizes. Without loss of generality this paper focuses on undirected unweighted graphs such that $(u, v) = (v, u)$ and $w(u, v) \in \{0, 1\}\ \forall u, v \in \mathcal{V}$ (1: link exists; 0: link does not exist; $w(u, u) = 0\ \forall u$). The simulation results are further compared in terms of Surprise optimality towards the set of graph clustering techniques integrated in the so-called SurpriseMe tool [19]. This tool sequentially executes a number of clustering techniques and ranks them in terms of their associated Surprise score. Therefore, for each network topology a different clustering technique can be selected and there exists no universal technique that outperforms any other over dintinct network topologies. Moreover, the goodness of the proposed local search method is envinced by comparing the obtained results with respect to the same technique without the application of the local search procedure. 10 MonteCarlo simulations and 300 iterations are executed by the proposed FA and FA+LS techniques with a TSR probability of 0.01.

Specifically, the techniques within the SurpriseMe tool are:

- CPM [20]: A Constant Potts Model.
- SCLUSTER [21]: A dendrogram-based approach aimed at performing iterative hierarchical cluster analysis.
- RNSC [22]: A Restricted Neighborhood Search Clustering algorithm.
- RN [23]: A local spin-glass-type Potts model for community detection that utilizes an absolute energy evaluation.
- INFOMAP [24]: An information-theoretic approach which relies on information flows sent over random walks through the graph so as to construct a map with enhanced information about the clustering structure of the graph.
- UVCLUSTER [25]: A hierarchical clustering scheme that build a dendrogram by means of iteratively explores distance datasets. The peer-to-peer distances between nodes are evaluated in order to compute the strength of the connection between nodes of a cluster.
- RB [26]: A graph clustering approach that infers the spin configuration that minimizes the energy of an infinite ranged Potts glass.

Furthermore, the so-called Louvain method often used to unveil community structures in networks by iterative aggregating nodes on a clustering hierarchy has been included in the benchmark as a reference (label LOUVAIN).

Table I depicts the (min/mean/max) values obtained by the proposed FA with Local Search method (FA+LS) and without the LS procedure (FA) over 6 network instances: Relaxed Caveman 50 nodes RC(50), Relaxed Caveman 100 nodes RC(100), Erdos Renyi 50 nodes ER(50), Erdos Renyi 100 nodes ER(100), Powerlaw Cluster Graph 50 nodes PL(50) and Powerlaw Cluster Graph 100 nodes PL(100):

- RC$(\theta, \vartheta, p_{rw})$ refers to graph instances generated by the Relaxed Caveman Model. This model starts with θ cliques of ϑ nodes. Such nodes are subsequently randomly rewired with probability p_{rw} to link different cliques. These parameters give rise to $|\mathcal{V}| = \theta \cdot \vartheta$ nodes grouped in *perfect* communities that become more loosely connected as p_{rw} increases.
- ER(η, p_{er}) denotes graph instances created by an Erdos-Renyi Model which consists of a random graph of η nodes with connections between them generated under probability p_{er}.
- PL$(\mu, \beta, p_{\triangle})$ refers to graph instances following the Power Law Cluster Graph Model. $|\mathcal{E}| = \mu$ nodes are progressively included to the graph. Per newly added

TABLE I
RANKING STATISTICS OF THE SURPRISE VALUES (*min/mean/max*) OBTAINED BY THE COMMUNITY DETECTION TECHNIQUES WHEN APPLIED TO 6 DIFFERENT GRAPH INSTANCES COMPUTED OVER 10 INDEPENDENT MONTE CARLO REALIZATIONS.

	FA+LS	FA	CPM	SCLUSTER	RNSC	RN	INFOMAP	UVCLUSTER	RB	LOUVAIN
RC(5,10,0.1)	3.36/129.56/184.07	59.25/116.83/193.48	193.48	193.48	193.48	193.48	193.48	193.48	193.48	193.48
RC(5,20,0.1)	234.97/690.63/798.93	11.53/599.97/764.36	798.93	798.93	798.93	798.93	798.93	798.93	798.93	798.93
ER(50,0.1)	10.50/29.10/37.88	5.55/18.84/37.78	36.33	38.14	31.71	37.31	33.02	35.97	34.97	25.56
ER(100,0.1)	21.47/80.33/89.09	28.17/74.49/88.30	85.12	75.99	81.86	3.40	0	86.01	83.21	57.69
PL(50,3,0.1)	21.08/28.89/31.74	8.36/23.33/31.84	28.47	29.21	29.84	30.48	0	29.68	25.52	21.64
PL(100,3,0.1)	15.73/64.75/80.22	11.97/59.22/78.47	76.70	69.69	78.67	71.24	64.79	70.58	68.43	51.71

TABLE II
MAXIMUM VALUES OF THE SURPRISE METRIC OBTAINED BY THE PROPOSED FA+LS ALGORITHM AND THE REST OF ALGORITHMS.

	FA+LS	CPM	SCLUSTER	RNSC	RN	INFOMAP	UVCLUSTER	RB	LOUVAIN
Dolphins	75.25	49.18	73.40	75.09	30.23	34.01	49.49	66.20	55.99

node β edges are created at random and a new edge connecting a given node to one of its neighbors named *triangle* is created with probability $p_{\triangle}$.

As can be observed by analyzing the results in Table I, mean values for Surprise metric are outperformed by using the proposed local search method in all the aforementioned network topologies. When analyzing the performance of Surprise maximization along the iterative process by FA and the proposed FA+LS technique, it can be observed in Figure 1 that by applying the local search procedure higher values for Surprise are obtained at lower iterations and a higher value is achieved during the iterative process. Moreover, a Wilcoxon test has been performed over FA and the proposed FA+LS. Results show that both distributions are statistically significant at a confidence level of 95% for RC and ER graph instances.

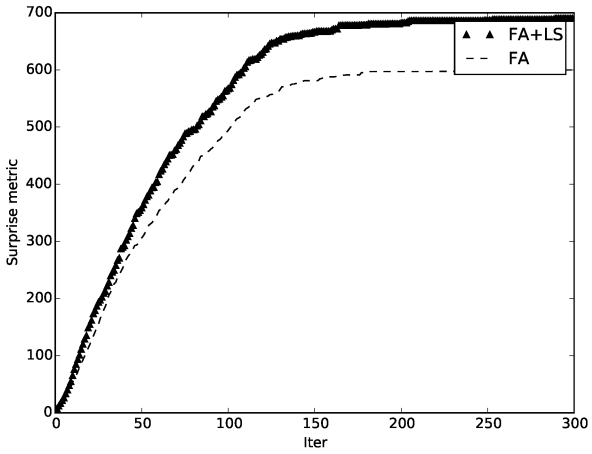

Fig. 1. Mean Surprise values along iterations for FA and FA+LS for the Relaxed Caveman network with 100 nodes RC(100).

When comparing maximum values of FA+LS with respect to results obtained by the rest of community detection algorithms embedded in the SurpriseMe tool, it can be noted that the FA+LS method is a universally competitive graph clustering proposal which behave well across networks of diverse topologies.

A. Real Case Study: Dolphins Network

Once the proposed algorithm has been validated in synthethic scenarios, in this subsection a real case scenario based on directed social network of bottlenose dolphins is presented. The nodes are the bottlenose dolphins (genus Tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand. An edge indicates a frequent association and the observations occurred between 1994 and 2001. As stated in [28] the network is composed of 62 dolphins and edges were set between animals that were seen together more often than expected by chance. The dolphins separated in two groups after a dolphin left the place for some time.

Table II depicts the maximum values obtained by the Surprise metric by the proposed FA+LS algorithm and the rest of community detection algorithms. As can be observed FA+LS algorithm is able to achieve the highest Surprise value which demonstrates its potential and suitability as a universal community detection technique.

V. CONCLUDING REMARKS

This papers has elaborated on a novel local search method for graph clustering based on population-based distribution learning capable of outperforming other community detection algorithms in the literature. The proposed scheme hinges on estimating the probability distribution of pairs of nodes over the set of cluster arrangements found by a global solver, from which new cluster proposals are sampled in a similar fashion to canonical estimation of distribution algorithms. In particular this work has described the incorporation of such a procedure to a Firefly algorithm specially devised for tackling graph clustering problems. Furthermore, the simplicity of the proposed method makes it a low-complexity alternative for its hybridization with other population-based heuristics.

The performance of the proposal has been tested over synthetic network topologies of different sizes, as well as over a real case scenario based on the Dolphins dataset. In general it has been shown that the proposed FA+LS technique is capable of achieving competitive results in a large number of networks while other proposals of SurpriseMe tool are well-suited for specific network topologies and do not behave well as an universal community detection tool.

Future research will be focused on extending this study to more diverse and real-based graphs, as well as undertaking the detection of overlapped communities. Efforts will be also conducted towards the parallelization of the search process of the algorithm in order to enhace its convergence.

ACKNOWLEDGMENTS

This work has been supported by the Basque Government through the ELKARTEK program (ref. KK-2015/0000080).

REFERENCES

[1] J. Chen, B. Yuan, "Detecting functional modules in the yeast proteinprotein interaction network", Bioinformatics, Vol. 22, N. 18, pp. 2283–2290, 2006.

[2] K. P. Reddy, M. Kitsuregawa, P. Sreekanth, S. S. Rao, "A graph based approach to extract a neighborhood customer community for collaborative filtering", Proceedings of the Second International Workshop on Databases in Networked Information Systems, pp. 188-200, Springer-Verlag, London, UK, 2002.

[3] M. Prateek, V. Vasudeva, "Improved topic models for social media via community detection using user interaction and content similarity", 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), pp. 1-7. IEEE, 2016.

[4] T. Chakraborty, A. Dalmia, A. Mukherjee, N. Ganguly, "Metrics for community analysis: A survey". arXiv preprint arXiv:1604.03512, 2016.

[5] J. Creusefond, T. Largillier, S. Peyronnet, "On the evaluation potential of quality functions in community detection for different contexts". In International Conference and School on Network Science, pp. 111-125, Springer International Publishing, 2016.

[6] U. Brandes, M. Gaertler, D. Wagner, "Experiments on Graph Clustering Algorithms", Proceedings of Algorithms/ESA 2003, Springer Lecture Notes in Computer Science, Vol. 2832, pp. 568–579, 2003.

[7] M. E. Newman, M. Girvan, "Finding and Evaluating Community Structure in Networks", Physical Review E, Vol. 69, N. 2, pp. 026113, 2004.

[8] R. Kannan, S. Vempala, A. Vetta, "On Clusterings: Good, Bad and Spectral", Journal of the ACM, Vol. 51, N. 3, pp. 497–515, 2004.

[9] R. Aldecoa, I. Marin, "Deciphering Network Community Structure by Surprise", PLoS ONE, Vol. 6(9), pp. e24195, 2011.

[10] R. Aldecoa, I. Marin, "Closed benchmarks for network community structure characterization", Phys. Rev. E 85, 026109, 2012.

[11] C. Nicolini, A. Bifone, "Modular structure of brain functional networks: breaking the resolution limit by Surprise". Scientific reports, 6, 2016.

[12] J. Del Ser, J. Lopez, E. Villar-Rodriguez, M. N. Bilbao, C. Perfecto, "Community Detection in Graphs based on Surprise Maximization using Firefly Heuristics", IEEE Congress on Evolutionary Computation (CEC), pp. 2233–2239, Vancouver, 2016.

[13] X. S. Yang, "Firefly Algorithms for Multimodal Optimization", Stochastic Algorithms: Foundations and Applications, pp. 169–178, Springer Berlin Heidelberg, 2009.

[14] A. Lancichinetti, S. Fortunato, "Consensus clustering in complex networks", Nature, Scientific Reports 2, Article number 336, 2012.

[15] T. Fleck, A. Kappes, D. Wagner, "Graph Clustering with Surprise: Complexity and Exact Solutions", SOFSEM 2014, Springer Lecture Notes in Computer Science, Vol. 8327, pp. 223–234, 2014.

[16] M. Tasgin, A. Herdagdelen, H. Bingol, "Community detection in complex networks using genetic algorithms", arXiv preprint arXiv:0711.0491, 2007.

[17] M. Chen, K. Kuzmin, B. K. Szymanski, "Community detection via Maximization of Modularity and Its Variants", IEEE Trans. Computation Social Systems, Vol. 1, N. 1, pp. 46–65, 2014.

[18] A. Strehl, J. Ghosh, "Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions", Journal of Machine Learning Research, Vol. 3, pp. 583–617, 2002.

[19] R. Aldecoa, I. Marin, "SurpriseMe: An Integrated Tool for Network Community Structure Characterization using Surprise Maximization", Bioinformatics, Vol. 30, N. 7, pp. 1041–1042, 2014.

[20] V. A. Traag, P. Van Dooren, Y. Nesterov, "Narrow Scope for Resolution-Limit-Free Community Detection", Physical Review E, Vol. 84, pp. 016114, 2011.

[21] R. Aldecoa, I. Marin, "Jerarca: Efficient Analysis of Complex Networks using Hierarchical Clustering", PLoS ONE, Vol. 5, pp. e11585, 2010.

[22] A. D. King, N. Przulj I. Jurisica, "Protein Complex Prediction via Cost-based Clustering", Bioinformatics, Vol. 20, pp. 3013, 2004.

[23] P. Ronhovde, Z. Nussinov, "Local Resolution-Limit-Free Potts Model for Community Detection", Physical Review E, Vol. 81, pp. 046114, 2010.

[24] M. Rosvall, C. T. Bergstrom, "Maps of Random Walks on Complex Networks Reveal Community Structure", Proceedings of the National Academy of Sciences of the USA, Vol. 105, pp. 1118, 2008.

[25] V. Arnau, S. Mars, I. Marin, "Iterative Cluster Analysis of Protein Interaction Data", Bioinformatics, Vol. 21, pp. 364, 2005.

[26] J. Reichardt, S. Bornholdt, "Statistical Mechanics of Community Detection", Physical Review E, Vol. 74, pp. 016110, 2006.

[27] P. Larrañaga, J. A. Lozano, eds., "Estimation of distribution algorithms: A new tool for evolutionary computation," Vol. 2, Springer Science & Business Media, 2001.

[28] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations", Behavioral Ecology and Sociobiology, Vol. 54, pp. 396–405, 2003.