Uncovering Communities in Multidimensional Networks with Multiobjective Genetic Algorithms

Alessia Amelio
National Research Council of Italy (CNR)
Inst. for High Perf. Comp. and Net. (ICAR)
Via P. Bucci 41C, 87036 Rende(CS), Italy
amelio@icar.cnr.it

Clara Pizzuti
National Research Council of Italy, CNR
Inst. for High Perf. Comp. and Net. (ICAR)
Via P. Bucci 41C, 87036 Rende(CS), Italy
pizzuti@icar.cnr.it

ABSTRACT

A framework for community discovery in multidimensional networks based on an evolutionary approach is proposed. Each network is clustered by running a multiobjective genetic algorithm that tries to maximize the modularity function of the current network and, at the same time, to minimize the difference between the current community structure and that obtained on the already considered dimensions. Experiments on synthetic datasets show the capability of the approach in discovering latent shared group organization of individuals.

Categories and Subject Descriptors

H.2.8 [**Database Managment**]: Database Applications — *Data Mining*; I.5.3 [**Computing Methodologies**]: Pattern Recognition—*Clustering*

General Terms

Algorithms

Keywords

Multi-dimensional Networks; Community detection; Multiobjective Genetic Algorithms

1. INTRODUCTION

In the last years, the rapid growth of social networking sites has provided people a media to communicate and exchange information. Each individual user often participates in different social networks with different strength, thus playing diverse roles. This implies that many real-world networks are multidimensional since actors are connected by different relationships. In the last years there has been an increasing interest in complex networks presenting multiple connections between pairs of individuals.

A multidimensional network can be viewed as a set of slice networks. Each slice, modeled as a graph, represents a facet of the individual activity, i.e. the connections among individuals in a particular dimension. Generally, the interactions of the same user may be rather different, since a user may be involved in distinct activities with variable concern. The objective in a multidimensional network is to uncover a shared community structure among objects such that a quality function is optimized in all the dimensions.

In this paper a new framework based on multiobjective optimization to deal with the problem of detecting a shared community structure in multidimensional networks is proposed. The competing objectives to optimize try to obtain a community structure for the i-th dimension as accurate as possible, and, at the same time, that does not differ too much from the clustering obtained so far on the i-1 already considered dimensions. Experiments on synthetic networks show that the multiobjective approach allows the detection of accurate community structures in multidimensional networks.

2. METHOD DESCRIPTION

A multidimensional network is a sequence $\mathcal{N} = \{\mathcal{N}_1, \ldots, \mathcal{N}_d\}$ of *slice networks*, where each $\mathcal{N}_l$ is a dimension and it is modeled as a graph $G_l = (V_l, E_l)$, being $V_l \subseteq V$ the set of nodes, and E_l the set of links connecting elements of V_l in the l-th dimension of $\mathcal{N}$. A clustering, or community structure, $\mathcal{CS}_l = \{C_1^l, \ldots C_k^l\}$ of a network $\mathcal{N}_l$ is a partitioning of G_l in groups of nodes that maximizes a quality function Q. Furthermore, for each couple of communities C_i^l and $C_j^l \in \mathcal{CS}_l$, $V_{l_i} \cap V_{l_j} = \emptyset$. Our objective is to uncover a shared community structure $\mathcal{CS}$ among the objects of the multidimensional network $\mathcal{N}$ such that the quality function Q is optimized in all the d dimensions.

The problem of finding a shared community structure in a multidimensional network can be viewed as the analogous problem in a dynamic network, i.e. a network that evolves by changing its interconnections over time, where each dimension of the multidimensional network corresponds to a time stamp in the dynamic network. In particular, the evolutionary clustering approach proposed by Chakrabarti et al. [1] and exploited in [2] for dynamic networks, can be extended to multidimensional networks by considering the concepts of *facet quality* $\mathcal{FQ}$ and *sharing cost* $\mathcal{SC}$ as the analogous of *snapshot quality* and *temporal smoothness* of evolutionary clustering. Facet quality guarantees that the clustering found for the i-th dimension under consideration maximizes a quality function as much as possible, while the sharing cost means that the clustering of the current facet agrees as much as possible with the clustering obtained for

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).
GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-1964-5/14/07.
http://dx.doi.org/10.1145/2598394.2598453.

the previously considered i-1 dimensions. In this new framework, given the sequence $A_1, \ldots, A_d$ of adjacency matrices associated with the graphs $\{G_1, \ldots, G_d\}$ modeling a multidimensional network $\mathcal{N} = \{\mathcal{N}_1, \ldots, \mathcal{N}_d\}$, a shared community structure among the networks $\mathcal{N}_i$ can then be obtained by iteratively optimizing both facet quality and sharing cost. The community structure obtained for the last dimension d thus can be considered the best sharing community structure among the d dimensions. A method to carry out this framework is to use a multiobjective genetic algorithm that finds a solution realizing the best trade-off between facet quality of network $\mathcal{N}_i$ and sharing cost with $\mathcal{N}_{i-1}$. As *facet quality* $\mathcal{FQ}$ we employ the well known concept of *modularity* introduced by Girvan and Newman [3], while the second objective is the *Normalized Mutual Information*, that gives the similarity between the community structure of the current facet and the clustering obtained so far for the facets already considered.

The multiobjective method consists of two main steps. In the former the network $\mathcal{N}_1$ is clustered by employing a genetic algorithm that optimizes the modularity value. In the second one, a multiobjective genetic algorithm, for each pair of dimensions $\mathcal{N}_i$ and $\mathcal{N}_{i-1}$, tries to optimize the facet quality $\mathcal{FQ}$ for the graph G_i modeling the current dimension $\mathcal{N}_i$, and the sharing cost $\mathcal{SC}$ computed as the normalized mutual information between the clustering obtained for G_i and that for G_{i-i}.

Both steps use the locus-based adjacency representation and uniform crossover. Moreover, the initialization process assigns to each node i one of its neighbors j, and the mutation operator randomly assigns to a node i one of its neighbors. The multiobjective genetic algorithm is iteratively executed for the d-1 dimensions by optimizing the two objectives $\mathcal{FQ}$ and $\mathcal{SC}$. For each iteration, the clustering having the best modularity value is chosen from the Pareto front as current solution. It is worth to note that the network ordering could influence the performances of the method because slices are considered sequentially, thus processing first one network instead of another could produce different results. Choosing the best ordering is not an easy task and deserves a deep investigation which is beyond the scope of this paper. Instead of choosing a random order, we employed a heuristic based on the the concept of clustering coefficient of a network. In the next section we show that the combination of multiobjective optimization and clustering coefficient based ordering gives good performance results, also compared with the Tang et al. methods [4] on synthetic networks.

3. EXPERIMENTAL RESULTS

In this section we present the results obtained by the method on synthetic data sets for which the ground-truth division in communities is known, and compare the method with the spectral-based approaches proposed by Tang et al. [4], besides the result obtained by a genetic algorithm optimizing modularity on a single dimension. The data set, proposed by Tang et al. [4], consists of 350 objects grouped into three clusters of 50, 100, and 200 objects, respectively. The number of dimensions is 4, i.e. the objects can interact in 4 different ways. An example can be seen in Figure 1. We executed the method on 50 different generated synthetic networks and computed the normalized mutual information NMI between the obtained clustering and the true community structure. As regards the parameters, we set population size 350, number of generations 200, crossover fraction 0.9, mutation rate 0.2. The implementation has been written in MATLAB 7.14 R2012a, using the Genetic Algorithms and Direct Search Toolbox 2.

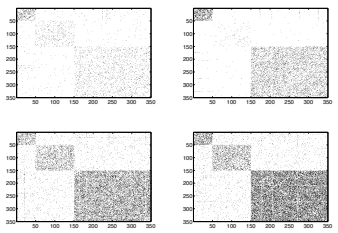

Figure 1: Example of 4-D synthetic network.

	Strategy	*Evolutionary*	*Spectral*
1-D	A1	0.7241 ± 0.2140	0.7237 ± 0.1924
	A2	0.8530 ± 0.0768	0.6798± 0.1888
	A3	0.7918 ± 0.1427	0.6672 ± 0.1848
	A4	0.8176 ± 0.0844	0.6906± 0.1976
4-D	*MultiMOGA*	0.9368 ± 0.0118	-
	PMM	-	0.9351 ± 0.1059
	AMM	-	0.7946 ± 0.1623
	TMM	-	0.9157 ± 0.1137

Table 1: Comparing the NMI values between the evolutionary computation approaches and spectral approaches of Tang et al. [4].

From Table 1 we can observe that the results obtained by *MultiMOGA* are superior with respect to the AMM and TMM methods. Regarding PMM, the NMI values of *MultiMOGA* are slightly higher, and our approach is more robust that PMM having a much lower standard deviation. On the single dimensional methods, the genetic approach always outperforms the spectral approach. It worth to point out that the spectral approaches need as input parameter the number of communities to find, while the GAs methods automatically determine this value because of the genetic representation that encodes the optimal division of objects with respect to the objective function to maximize.

4. REFERENCES

[1] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (KDD'06)*, pages 554–560, 2006.

[2] F. Folino and C. Pizzuti. An evolutionary multiobjective approach for community discovery in dynamic networks. *IEEE Transactions on Knowledge and Data Engineering, to appear*, 2014.

[3] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E69:026113, 2004.

[4] L. Tang, X. Wang, and H. Liu. Uncoverning groups via heterogeneous interaction analysis. In *The Ninth IEEE International Conference on Data Mining (ICDM'09)*, pages 503–512, 2009.