Finding Redundant and Complementary Communities in Multidimensional Networks

Michele Berlingerio
KDDLab ISTI-CNR
Via G. Moruzzi, 1, 56124
Pisa, Italy
michele.berlingerio@isti.cnr.it

Michele Coscia
KDDLab University of Pisa
Largo B. Pontecorvo, 3, 56127
Pisa, Italy
coscia@di.unipi.it

Fosca Giannotti
KDDLab ISTI-CNR
Via G. Moruzzi, 1, 56124
Pisa, Italy
fosca.giannotti@isti.cnr.it

ABSTRACT

Community Discovery in networks is the problem of detecting, for each node, its membership to one of more groups of nodes, the *communities*, that are densely connected, or highly interactive. We define this problem for multidimensional networks, i.e. where more than one connection may reside between any two nodes. We introduce two measures able to characterize the communities found. Our experiments on real world data support the methodology proposed, and open the way for a new class of algorithms, aimed at capturing the multifaceted complexity of connections among nodes in a network.

Categories and Subject Descriptors: G.2.2 [Graph Theory]: Graph Algorithms; H.2.8 [Database Applications]: Data Mining

General Terms: Algorithms, Theory

Keywords: Multidimensional Network Analysis, Community Discovery

1. INTRODUCTION

Inspired by real-world scenarios such as social networks, technology networks, the Web, and so on, in the last years, wide and multidisciplinary research has been devoted to the extraction of non trivial knowledge from such networks. One crucial task in network analysis is Community Discovery, i.e., the discovery of group of nodes densely connected, or highly related. There exist many techniques able to identify communities in networks [2], allowing to detect hierarchical connections, influential nodes in communities, or just group of nodes that share some properties or behaviors. Among the most popular approaches, we recall: the prolific modularity-oriented class [1]; a label propagation approach [5]; and a community discovery algorithm based on random walks [4]. Most of the existing approaches are limited to monodimensional networks, i.e. networks where there can be only one interaction between any two nodes. We deal with *multidimensional networks*, where multiple connections may exist between a pair of nodes, reflecting various interactions (i.e., dimensions) between them. Multidimensionality in real networks may be expressed by either different types of connections (two persons may be connected because they are friends, colleagues, and so on), or different quantitative values of one specific relation (co-authorship between two authors may occur in several different years, for example). In this scenario, we introduce the problem of *Multidimensional Community Discovery.* An example of multidimensional community discovery algorithm exists in literature. In [3] the authors extend the definition of modularity to fit to the multidimensional case, which they call “multislice”. However, in this work authors do not consider any definition of “multidimensional community”, neither they characterize and analyze the communities found.Instead, we define a concept of multidimensional community, and we introduce two new measures aimed at analyzing the multidimensional properties of the communities discovered. We then show the results obtained by applying our framework on real-world networks, giving a few examples of interesting multidimensional communities found in a movie collaboration network.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

2. FINDING AND CHARACTERIZING MULTIDIMENSIONAL COMMUNITIES

We use a *multigraph* to model a multidimensional network and its properties. For the sake of simplicity, in our model we only consider undirected multigraphs and since we do not consider node labels, hereafter we use *edge-labeled undirected multigraphs*, denoted by a triple $\mathcal{G} = (V, E, D)$ where: V is a set of nodes; D is a set of labels; E is a set of labeled edges, i.e. the set of triples (u, v, d) where $u, v \in V$ are nodes and $d \in D$ is a label. Also, we use the term *dimension* to indicate *label*, and we say that a node *belongs to* or *appears in* a given dimension d if there is at least one edge labeled with d adjacent to it. We assume that given a pair of nodes $u, v \in V$ and a label $d \in D$ only one edge (u, v, d) may exist. Thus, each pair of nodes in $\mathcal{G}$ can be connected by at most $|D|$ possible edges.

2.1 Multidimensional Community

Adding multidimensionality to the problem of community discovery leads to an opinable concept of multidimensional community. We start with a high-level possible definition, then we add more semantic to it.

Definition 1 (Multidimensional Community). *A multidimensional community is a set of nodes densely connected in a multidimensional network.*

While in a monodimensional network the density of a com-

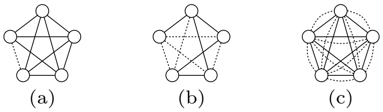


Figure 1: Multidimensional communities

munity refers unambiguously to the ratio between the number of edges among the nodes and the number of all possible edges, the multidimensional setting offers an additional degree of freedom. Consider Figure 1: in (a) we have a community whose density mostly depends by the connectivity provided by one dimension; in (b) we have a different situation, as both the dimensions are contributing to the density of the community. Should the two be considered equivalent or can we discern among them? In order to answer this question, we define two measures, γ and ρ, aimed at capturing two different phenomena that can be detected in a community. Hereafter, we use this notation: c is a multidimensional community; d is a dimension in D; D_c is the subset of D appearing in c; P is set of pairs (u, v) connected by at least one dimension in the network; $\overline{P} \subseteq P$ is the set of pairs (u, v) connected exclusively by one dimension; $\overline{\overline{P}} = P \setminus \overline{P}$ is the set of pairs connected by at least two dimensions; P_c is the subset of P appearing in c; $P_{c,d}$ is the set of pairs (u, v) in c connected at least in d and $\overline{P_{c,d}} \subseteq P_{c,d}$ is the set of pairs (u, v) in c connected exclusively in d; $\overline{\overline{P_c}} \subseteq \overline{\overline{P}}$ is the subset of $\overline{\overline{P}}$ containing only pairs in c.

2.2 Complementarity γ

The first measure, γ, that we call *complementarity*, is the conjunction of three concepts: **variety** $\mathcal{V}_c$, i.e. how many different dimensions are detectable among the community c; **exclusivity** $\mathcal{E}_c$, i.e. how many pairs of nodes are connected by just one dimension within c; **homogeneity** $\mathcal{H}_c$, i.e. how uniform is the distribution of the number of edges per dimension in c. We want this measure to be higher when each of the above is high. A natural way to achieve this is to aggregate them by their product:

$$\gamma_c = \mathcal{V}_c \times \mathcal{E}_c \times \mathcal{H}_c \tag{1}$$

We now have to define the three concepts. Variety can be computed by

$$\mathcal{V}_c = \frac{|D_c| - 1}{|D| - 1} \tag{2}$$

as the number of dimensions expressed with the community c over the total number of dimensions within the network. The two negative terms serve as corrections to make Variety take values in $[0, 1]$. Note that Variety defined as above would be undefined when $|D| = 1$, but this would mean having a monodimensional network, where the use of γ would be meaningless.

Exclusivity can be computed as the ratio between the number of exclusive connections within the community and the total number of connected pairs in c:

$$\mathcal{E}_c = \frac{\sum_{d \in D} |\overline{P_{c,d}}|}{|P_c|} \tag{3}$$

This term is equal to zero when there are no exclusive connections, i.e. every pair of nodes in c is connected by at least two dimensions, while it is equal to one when every pair in c is connected by only one dimension. The formula is not defined for $|P_c| = 0$, which happens only for communities of only one node, for which it has no sense to compute γ.

Finally, we have to define Homogeneity. We want this term to be equal to one when the edges within the community are uniformly distributed among the dimensions represented in c. The simplest way to measure this is to look at the standard deviation of the distribution of the edges in c on the dimensions. We define:

$$\sigma_c = \sqrt{\frac{\sum_{d \in D} (|P_{c,d}| - avg_c)^2}{|D|}} \tag{4}$$

where avg_c is the mean of the distribution, as the standard deviation of the number of edges per dimension in c, and:

$$\sigma_c^{max} = \sqrt{\frac{(max(|P_{c,d}|) - 1)^2}{2}} \tag{5}$$

where $max(|P_{c,d}|)$ is the number of edges belonging to the dimension more represented in c, as the maximum theoretic standard deviation. Then, we can define Homogeneity as:

$$\mathcal{H}_c = 1 - \frac{\sigma_c}{\sigma_c^{max}} \tag{6}$$

where we subtract the right term to one, in order to make $\mathcal{H}_c$ equal to one when the right term is zero, i.e. when the edges are uniformly distributed among the different dimensions.

If we could have the complete set of communities of a network, we could make a more precise estimation of σ_c^{max}:

$$\sigma_c^{max} = \sqrt{\frac{(max(|P_{c,d}|) - min(|P_{c,d}|))^2}{2}} \tag{7}$$

where $min(|P_{c,d}|)$ is the number of edges belonging to the dimension having the lowest number of edges among all the communities.

In the exceptional case in which all the communities would see all the dimensions represented with the same number of edges, the two normalization coefficients σ_c^{max} would be equal to zero, making the right term of Equation 3 undefined. In this case, being the denominator an upper bound, also the numerator would be equal to zero. But this is the ideal topology of a network where the Homogeneity is maximum since all the edges are uniformly distributed, and then we can handle this case, without lack of generality, by defining $\mathcal{H}_c$ as:

$$\mathcal{H}_c = \begin{cases} 1 & \text{if } \sigma_c = 0 \\ 1 - \dfrac{\sigma_c}{\sigma_c^{max}} & \text{otherwise} \end{cases} \tag{8}$$

Example 1 (multidimensional communities and γ). *Consider Figure 1. We see three different multidimensional communities, each of them with different multidimensional structures: in (a), the standard deviation of the number of edges per dimension is the maximum possible, hence $\mathcal{H}_c = 0$, thus $\gamma = 0$; in (b), every term of the complementarity is equal to one, thus $\gamma = 1$; in (c), the exclusivity is zero, as every pair is connected by two dimensions, hence $\gamma = 0$.*

2.3 Redundancy ρ

The second measure we define is the *redundancy*, capturing the phenomenon for which a set of nodes that constitute a community in a dimension, constitute a community also in other dimensions. We can see this measure as a simple indicator of the redundancy of the connections: the more dimensions connect each pair of nodes within a community, the higher the redundancy will be. We can then define ρ by counting how many pairs have redundant connections, normalizing by the theoretical maximum:

$$\rho_c = \sum_{(u,v) \in \overline{\overline{P_c}}} \frac{|\{d : \exists (u, v, d) \in E\}|}{|D| \times |P_c|} \tag{9}$$

With the help of Figure 1 we see how ρ takes values in $[0, 1]$: in 1(b), each pair of nodes is connected in only one dimension, then $|\overline{\overline{\overline{P_c}}}| = 0$ and the numerator is equal to zero; in 1(c), all the node pairs are connected in all the dimensions of D, which is equivalent to the number of connected pairs $|P_c|$ multiplied by the number of network dimensions $|D|$ (the denominator), making $\rho = 1$. We see that ρ is undefined for communities formed by one single node, where $|P_c| = 0$ and then the denominator is equal to zero. For this type of communities, however, the redundancy measure is not meaningful, thus we can ignore this case.

2.4 Problem definition

We can now formulate the problem under investigation:

PROBLEM 1 ($\mathcal{MCD}$). *Given a multidimensional network $\mathcal{G}$, find the complete set of multidimensional communities $\mathcal{C}$, and characterize each $c \in \mathcal{C}$ according to γ and ρ.*

3. A FRAMEWORK FOR $\mathcal{MCD}$

A complete solution for our problem would require to design an algorithm for extracting multidimensional communities, driven by the multidimensional density of the connections among nodes. However, it is difficult to define multidimensional density as universal, which is exactly what makes γ and ρ both meaningful. In addition, we believe that trivial design choices may lead to an algorithm producing communities with distributions of γ and ρ possibly unfairly unbalanced by the decisions taken. Moreover, we believe that the main contributions of this paper are the problem definition and the characterization of the communities by the introduction of γ and ρ. For all these reasons, here we propose a different solution based on existing, monodimensional, algorithms.

In order to be able to apply existing solutions to multidimensional network, and to be able to extract multidimensional communities, we have to introduce a mapping function ϕ able to transform a multidimensional network in a monodimensional one, trying to keep as much information as possible, and a function ϕ' which recovers multidimensional information from monodimensional communities. The logical workflow to solve $\mathcal{MCD}$ is then:

$$\mathcal{G} \xrightarrow{\phi} G \xrightarrow{CD} C \xrightarrow{\phi'} \mathcal{C} \rightarrow \gamma, \rho\ (c \in \mathcal{C}) \tag{10}$$

where ϕ is a function that converts a multidimensional network $\mathcal{G}$ to a monodimensional network G, CD is any algorithm for community discovery on monodimensional networks, ϕ' is a function that, for each monodimensional community c, restores the multidimensional connections originally residing among the nodes of c in $\mathcal{G}$, thus returning a set of multidimensional communities $\mathcal{C}$, on which we are then able to compute our evaluating functions γ and ρ.

3.1 Three possible ϕ mappings

There can be several different definitions for ϕ, leading to different monodimensional networks built from $\mathcal{G}$. One possible class of them can be designed by simply *flattening* multidimensional edges to monodimensional ones, possibly weighting the monodimensional edges by some functions of the original multidimensional structure. In the following we assume to use a weight-based class of ϕ functions, and we define three different weighting strategies ϕ.

The first we define is μ and requires to weight the (u, v) edge in G with 1 if there exists at least one dimension connecting u and v in $\mathcal{G}$, or, in formula:

$$\mu_{u,v} = \begin{cases} 1 & \text{if } \{\exists\ d : (u, v, d) \in E\} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

In the remainder of the paper, we refer is as ϕ_μ. This flattening clearly looses most of the multidimensional information residing in $\mathcal{G}$, except the neighborhood: any two nodes connected in $\mathcal{G}$ are also connected in G. One small improvement would be counting the number of dimensions connecting any two nodes u and v and using this as weight for the monodimensional edge added. We call this weight ν, which can be defined as:

$$\nu_{u,v} = |\{d : (u, v, d) \in E\}| \tag{12}$$

and we refer to the ϕ built upon ν as ϕ_ν.

We now consider a slight modification of ν that, instead of taking into account only the connection between u and v, also looks at their neighborhood, motivated by the intuition that common neighbors will likely be in the same community of u and v. We refer to this weight as η and define it as:

$$\eta_{u,v} = 1 + \frac{|N_{u,l} \cap N_{v,l}|}{|N_{u,l} \cup N_{v,l}| - 2} \tag{13}$$

where $N_{\cdot,l}$ is the set of neighbors in dimension d for a node. This is actually a multidimensional version of the clustering coefficient, and, according to the intuition behind it, should be able to better reflect the strength of the ties. We call ϕ_η the weight based on η. Note that there could be many other possible weighting strategy, as well as other different class of ϕ relying on different principles.However, to keep complexity low, and for sake of simplicity, in this paper we only examine the results obtained by using the three ϕ defined above.

3.2 The choice for CD

Any algorithm for community discovery can be used in our workflow, with one *caveat*: it should be able to handle edge weights. In our experiments, we present the results obtained by using an algorithm based on random walk [4], one based on label propagation [5] and one based of the fast greedy optimization of the modularity [1] as choices for possible monodimensional community discoverer. In our analysis we show how the choice among these three does not significantly affect the resulting distribution of γ and ρ.

3.3 Returning multidimensional communities

To get back restoring the original multidimensional information for each connected pair we only have to restore its original multidimensional connectivity in $\mathcal{G}$.

4. EXPERIMENTS

Our network was extracted from the Internet Movie Database (http://www.imdb.com). It is a collaboration network of years 2000-2009, where each node represents a person involved in a movie, and two persons are connected if they where involved in the same movie. We considered each year as a dimension of the network. Basic statistics of these networks are reported in Table 1. The framework, available for download[1], was implemented using R and igraph.

For the CD step, we chose three different algorithms: on based on random walk [4] (WT), one based on label propagation [5] (LP) and one based of the fast greedy optimization of the modularity [1] (FGM). WT and FGM returns the complete dendrograms of the communities, thus chose to take the cut maximizing the modularity as the best cut.

[1] http://kdd.isti.cnr.it/MCDF

Network	$\|V\|$	$\|E\|$	$\|P\|$	$\|D\|$	k	N	#cc	%GC	%SE
IMDb	28042	1291625	1131951	10	92.12	80.73	28	99.77	79.13

Table 1: Basic statistics: k is the average degree, N the average num. of neighbors, $\#cc$ the num. of connected components, $\%GC$ the ratio of nodes in the giant component, $\%SE$ is computed as $|\overline{P}|/|E|$

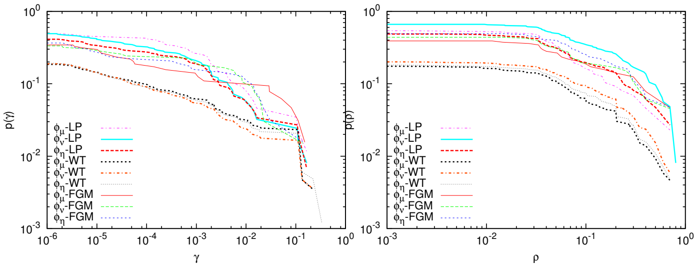


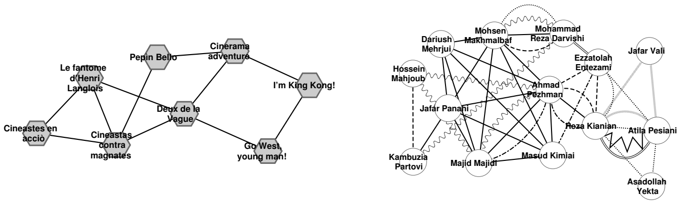
Figure 2: The cumulative distributions for γ and ρ in IMDb (color image).

Network	ϕ	LP		WT		FGM	
		$\|\mathcal{C}\|$	Q	$\|\mathcal{C}\|$	Q	$\|\mathcal{C}\|$	Q
IMDb	ϕ_μ	87	0.415	860	0.494	64	0.442
	ϕ_ν	124	**0.483**	847	**0.541**	66	**0.536**
	ϕ_η	148	0.460	823	0.507	63	0.530

Table 2: Communities ($|\mathcal{C}|$) and modularity (Q)

4.1 Quantitative Evaluation

Purpose of this section is to give a quantitative analysis of the results obtained driven by the following questions: **Q1** Can we evaluate the performances of the different choices of ϕ and CD? **Q2** How does the choice of ϕ and CD affect the distribution of γ and ρ over the communities? **Q3** What is the best choice of ϕ and CD parameters? In order to answer Q1, we looked at the values of the modularity measure (as defined in [1]), computed on the resulting set of communities C. This measure gives a value between -1 and 1, indicating how "good" nodes where partitioned into groups. The higher the value of modularity, the higher the partitioning reflects the division in the community of the graph that maximizes intra-community edges and minimizes inter-community edges. In Table 2 we report the modularity values, highlighting in bold, for each algorithm, which ϕ produced the highest value. We are interested in seeing whether a specific combination of ϕ and CD tends to produce higher scores. We notice that according best preprocessor was ϕ_ν , as it produces the best partition with each algorithm. In order to answer Q2, we analyzed the distribution of γ and ρ for the output of each ϕ-algorithm combination. These distributions are depicted in Figure 2. The distributions are generally overlapping and there is not a universally dominant combination. This confirms that our workflow does not significantly affect the distribution of the two measures. In addition, the information in Figure 2 may be used in conjunction with modularity in order to achieve richer knowledge about the results. Modularity, in fact, indicates how well the network is partitioned, and γ and/or ρ distributions characterize the multidimensional structure of the partitioning. One last consideration can be done looking at Table 2: there is no strong prevalence of one choice of parameters over the others. The same happens also for the distributions of γ and ρ. This suggest that the best answer for Q3 really depends on the final analysis of the network: the application scenario, the semantic of the dimensions and the time budget for running the experiments should drive the analyst towards the proper choice of the two parameters for our framework.



(a) high γ (0.005) in IMDb

(b) high ρ (0.05) in IMDb

Figure 3: A few interesting communities found, with their γ or ρ.

4.2 Analysis of Interesting Communities

We extracted two examples of communities with a relatively high (among the top 10%) score of γ and one with a relatively high ρ. We found one community with high complementarity, which was too large to be easily represented (more than 150 nodes). In Figure 3(a) we give a representation of it: each node is a clique within the community. Each group was found to represent a different documentary (titles provided as node labels) dedicated to a few important persons related to the cinema. The community has high complementarity because the personalities in a single documentary (such as Alfred Hitchcock or Jean-Luc Godard) are connected only by the year of release of the documentary itself, because they were not in activity in our sample of years, therefore not bound together by their own works. The connections between the documentaries are due to some movie stars present in both films. Redundancy in IMDb is able to identify large teams with continuous collaborations along many years. One group following this rule is composed by some masters of the Iranian cinema (Figure 3(b)), like Makhmalbaf and Panahi, both prominent directors in the Cannes and Venice Festivals.

5. CONCLUSIONS AND FUTURE WORK

We have addressed the problem of finding and characterizing communities in multidimensinal networks. We have given a possible definition of multidimensional community and then provided two different measures aimed at quantify and disambiguate the *density* of the community found, and devised a framework for our problem. Our results obtained on real world networks are encouraging, and provided a basis for future research on this direction. We plan to investigate the possibility of creating a multidimensional community discovery algorithm driven by γ and ρ scores.

6. REFERENCES

[1] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[3] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science 328, 876*, 2010.

[4] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *ISCIS 2005*, volume 3733, chapter 31, pages 284–293. Springer, Berlin, Heidelberg, 2005.

[5] Usha Nandini Raghavan, Reka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76:036106, 2007.