2011 International Conference on Advances in Social Networks Analysis and Mining

Finding and Characterizing Communities in Multidimensional Networks

Michele Berlingerio [1] Michele Coscia [2] Fosca Giannotti[1]

[1]KDDLab ISTI-CNR
Via G. Moruzzi, 1, 56124 Pisa - Italy
{michele.berlingerio, fosca.giannotti}@isti.cnr.it
Telephone: +39 0503152999
Fax: +39 0503152040

[2]KDDLab University of Pisa
Largo B. Pontecorvo, 3, 56127 Pisa - Italy
coscia@di.unipi.it
Telephone: +39 0502212752
Fax: +39 0502212726

***Abstract*—Complex networks have been receiving increasing attention by the scientific community, also due to the availability of massive network data from diverse domains. One problem studied so far in complex network analysis is Community Discovery, i.e. the detection of group of nodes densely connected, or highly related. However, one aspect of such networks has been disregarded so far: real networks are often multidimensional, i.e. many connections may reside between any two nodes, either to reflect different kinds of relationships, or to connect nodes by different values of the same type of tie. In this context, the problem of Community Discovery has to be redefined, taking into account multidimensionality. In this paper, we attempt to do so, by defining the problem in the multidimensional context, and by introducing also a new measure able to characterize the communities found. We then provide a complete framework for finding and characterizing multidimensional communities. Our experiments on real world multidimensional networks support the methodology proposed in this paper, and open the way for a new class of algorithms, aimed at capturing the multifaceted complexity of connections among nodes in a network.**

I. Introduction

Inspired by real-world scenarios such as social networks, technology networks, the Web, biological networks, and so on, in the last years, wide, multidisciplinary, and extensive research has been devoted to the extraction of non trivial knowledge from such networks. Predicting future links among the actors of a network ([13], [4]), detecting and studying the diffusion of information among them ([3]), mining frequent patterns of users' behaviors ([2], [20], [7]), are only a few examples of the objective in the field of Complex Network Analysis, that includes, among all, physicians, mathematicians, computer scientists, sociologists, economists and biologists.

The data at the basis of this field of research is huge, heterogeneous, and semantically rich, and this allows to identify many properties and behaviors of the actors involved in a network. One crucial task at the basis of Complex 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 ([11], [9]), allowing to detect hierarchical connections, influential nodes in communities, or just group of nodes that share some properties or behaviors. In order to do so, the connections among the nodes of a network are posed at the center of investigation, since they play a key role in the study of the network structure, evolution, and behavior.

Nowadays, most of the work done in the literature is limited to a very simplified perspective of such relations, focusing only on whether two nodes are connected or not. In the real world, however, this is not always enough to model all the available information, especially if the actors are users, with their multiple preferences, their multifaceted behaviors, and their complex interactions. A more sophisticated analysis of these element would help all the techniques basing their efficacy on the knowledge of the structure of a network.

To this aim, in this paper we deal with *multidimensional networks*, i.e. networks in which 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, they play together in a team, and so on), or different quantitative values of one specific relation (co-authorship between two authors may occur in several different years, for example). We can also distinguish between *explicit* or *implicit* dimensions, the former being relationships explicitly set by the nodes (friendship, for example), while the latter being relationships inferred by the analyst, that may link two nodes according to their similarity or other principles (two users may be passively linked if they wrote a post on the same topic).

In this scenario, we introduce the problem of *Multidimensional Community Discovery*, i.e. the problem of detecting communities of actors in multidimensional networks. We define a concept of multidimensional community, and we introduce a new measure aimed at analyzing the multidimensional properties of the communities discovered. We then present a framework for finding and characterizing multidimensional communities and we show the results obtained by applying such framework on real-world networks, giving a few examples of interesting multidimensional communities found in different scenarios: movie collaborations and terrorist attacks.

Our main contribution is then: we introduce and formally define the problem of multidimensional community discovery; we introduce a measure for characterizing the communities

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.104

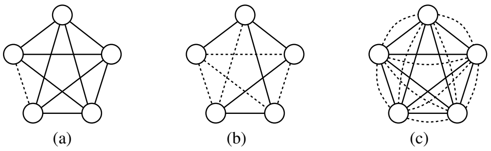

Fig. 1. Three examples of multidimensional communities

found; we build up a framework for solving the introduced problem by means of a conjunction of existing techniques and our newly introduced concepts; we perform a case study on real networks, showing the results obtained and the characterization of the communities.

II. Finding and characterizing multidimensional communities

In this section, after a model for multidimensional networks, we define multidimensional communities, a measure aimed at characterizing them, and the problem treated in this paper.

A. A model for multidimensional networks

We use a *multigraph* to model a multidimensional networks 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 also say that an edge *belongs to* or *appears in* a dimension d if its label is d. 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.

B. Multidimensional Community

As we see in Section III, the literature on community discovery presents a large number of diverse definitions of community. Adding multidimensionality to the problem leads to an even more opinable concept of multidimensional community. We start with a high-level possible definition, then we try to add more semantic to it.

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

As we see, while in a monodimensional network the density of a community 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 (i.e., the different dimensions). 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 a measure, ρ, aimed at characterizing multidimensional communities. In order to make it possible to compare its values among different networks, we make it take values in $[0, 1]$. In the following we use this notation: c is a multidimensional community; d is a dimension in D; P is set of pairs (u, v) connected by at least one dimension in the network; $\overline{\overline{P}}$ is the set of pairs connected by at least two dimensions P_c is the subset of P appearing in c; $\overline{\overline{P_c}} \subseteq \overline{\overline{P}}$ is the subset of $\overline{\overline{P}}$ containing only pairs in c.

C. Redundancy ρ

The measure we define is called *redundancy*, and it captures the phenomenon for which a set of nodes that constitute a community in a dimension tend to 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 the redundancy ρ by counting how many pairs have redundant connections, normalizing by the theoretical maximum:

$$\rho_c = \sum_{(u,v)\in\overline{\overline{P_c}}} \frac{|\{d : \exists(u, v, d) \in E\}|}{|D| \times |P_c|} \qquad (1)$$

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 $|\overline{\overline{P_c}}| = 0$ 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.

D. 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 ρ.

III. Related work

There are many studies on community detection, in various fields of research: computer science, physics, sociology, and others. Most of them can be grouped according to the definition of community they use.

One possibility is defining a community as a set of nodes with a high density of links among them, while there are sparse connections among different communities. The papers working with this definition rely on information theoric principles [15] or on the notion of modularity [6], which if a function defined to detect the ratio between intra- and inter-community number of edges. Modularity is widely studied and extended in many works: one of them is a greedy optimization able to scale up to networks with billions of edges [5].

Other works rely on some statistical properties of the graph. In [10], a framework for the detection of overlapping communities, i.e. communities allowing the vertices to be in more than one community, is presented.

Another class of approaches rely on the propagation in the network of a label [17] or a particular definition of structure (usually a clique [14]). The first approach is known for being a quasi linear solution for the problem, the second one allows to find overlapping communities.

One algorithm that tries to maximize quality and quantity measures on its results is InfoMap [18], a random walk-based algorithm. An emerging novel problem definition can be found in [1], in which authors state that community discovery algorithms should not group nodes but edges, emphasizing the role of the relation residing in a community.

Since 2009, multidimensionality has started to be taken into account in the community discovery problem. To the best of our knowledge, the main approaches are two. In [12] the authors extend the definition of modularity to fit to the multidimensional case, which they call "multislice". In [19] the authors create a machine learning procedure which detects the possible different latent dimensions among the entities in the network and uses them as features for the classification algorithm. It is important to note that both approaches do not consider any definition of "multidimensional community", neither they characterize and analyze the communities found and their multidimensional structure: their main limitation is to simply define a method for dealing with multidimensional networks, extracting monodimensional communities as output.

IV. A SOLUTION FOR $\mathcal{MCD}$

Given the problem definition above, a complete solution for it would require to design and develop an algorithm for extracting multidimensional communities, driven by the multidimensional density of the connections among nodes. However, according to our vision, it is difficult to define multidimensional density as universal, which is exactly what makes ρ meaningful. In addition, we believe that trivial design choices may lead to an algorithm producing communities with a distribution of 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 ρ. For all these reasons, we leave for future research the design and implementation of a multidimensional community discoverer able to exploit the additional degree of freedom that multidimensionality provides, and 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, whose function is 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:

$$\mathcal{G} \xrightarrow{\phi} G \xrightarrow{CD} C \xrightarrow{\phi'} \mathcal{C} \rightarrow \rho \ (c \in \mathcal{C}) \tag{2}$$

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 function and ρ.

We next give possible definitions of ϕ, we discuss which algorithm to use as CD, and we see how to implement ϕ'.

A. 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. An observation in support for this strategy is that many community discoverer use edge weights to reflect a more sophisticated definition of *dense* connections. In the following we assume to use a weight-based class of ϕ functions, and, in order to try to preserve as much multidimensional information as possible, we define three different weighting strategies, leading to three different ϕ.

The first weight 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:

$$\mu_{u,v} = \begin{cases} 1 & \text{if } \{\exists\ d : (u, v, d) \in E\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the remainder of the paper, we refer to the ϕ designed with this weight 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.

Can we do better? Can we preserve more of the original information? 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:

$$\nu_{u,v} = |\{d : (u, v, d) \in E\}| \tag{4}$$

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:

$$\eta_{u,v} = \frac{|N_{u,l} \cap N_{v,l}|}{|N_{u,l} \cup N_{v,l}| - 2} \tag{5}$$

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.

Note that there could be many other possible weighting strategy, as well as other different class of ϕ relying on different principles. For example, one might considering using the betweenness centrality instead of the clustering coefficient, or it is possible to consider also even more sophisticated measures. Note, however, that this could also mean additional computational complexity at the pre-processing stage, that adds to the community discoverer to be used afterward. 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.

B. The choice for CD

At this stage, any algorithm for community discovery can be used, with one *caveat*: we built a class of weight-based ϕ functions. This has to be taken into account by the algorithm, thus the only limitation we pose is to choose an algorithm able to handle edge weights. In our experiments, we present the results obtained by using an algorithm based on random walk [16], one based on label propagation [17] and one based of the fast greedy optimization of the modularity [6] as choices for possible monodimensional community discoverer. In our analysis we show how the choice among these three does not considerably affect the resulting distribution of ρ.

C. Returning multidimensional communities via ϕ'

Last question remained open so far about our workflow is, given the set of monodimensional communities returned by the CD step, how to get back restoring the original multidimensional information. This step turns out to be trivial, as, for every community, we have the set of the IDs of the nodes involved, and we can easily connect them with the original edges connecting them in $\mathcal{G}$.

V. Experiments

In order to validate our framework, we tested it on different real world networks, extracted from various sources. In this section, we provide the results obtained in this phase.

A. Tools, algorithms and running times

We ran our experiments on a server with 2 Xeon processors at 3.2GHz, 8GB of RAM, running Linux. The framework was implemented using R, making use of the igraph[1] library.

For the CD step, as stated above, we chose three different algorithms: on based on random walk [16], one based on label propagation [17] and one based of the fast greedy optimization of the modularity [6]. In the rest of the paper we refer to them as WT, LP and FGM. Note that, while LP and FGM are parameter-free, WT requires the length of the walk (that we set to 4 after empirical observations). Note also that WT and FGM returns the complete dendrograms of the communities, thus we had to choose a way to cut it. We then decided to take the cut maximizing the modularity as the best cut.

Given that the most computational expensive step in our framework is the extraction of monodimensional communities by external algorithms, we do not provide an extensive study of the running times. However, a single network took at most five hours to be processed by the nine combinations of pre-processors(ϕ) and algorithms for CD. We also report that the running time for computing ϕ, ϕ', and ρ are marginally relevant on the total running time.

Network	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert P\rvert$	$\lvert D\rvert$	k	N	#cc	%GC	%SE
GTD	2509	25200	24267	124	20.08	19.34	46	95.53	85.98
IMDb	28042	1291625	1131951	10	92.12	80.73	28	99.77	79.13

TABLE I
Statistics of the networks: k is the avg degree, N the avg number of neighbors, $\#cc$ the number of components, $\%GC$ the percentage of nodes in the Giant Component, $\%SE$ is the number of single edges connecting a pair

Network	ϕ	LP		WT		FGM	
		$\lvert\mathcal{C}\rvert$	Q	$\lvert\mathcal{C}\rvert$	Q	$\lvert\mathcal{C}\rvert$	Q
GTD	ϕ_μ	122	**0.622**	192	0.620	74	0.584
	ϕ_ν	109	0.547	197	0.603	65	0.611
	ϕ_η	165	0.500	194	**0.621**	78	**0.616**
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 II
Number of communities found ($|\mathcal{C}|$) and modularity (Q) for each combination of network, ϕ and algorithm.

B. Networks

For our study, we created the following multidimensional networks from real world data:

- **GTD**: From the database of global terrorism[2], we created a group-group network in which each terrorist organization is connected to another one if they have performed an attack in the same country, in the same year. The dimensions of this network are defined as the attacked country. In the orginal GTD database, the records include roughly 2k organizations (our nodes) active in 124 countries (our dimensions).
- **IMDb**: From the Internet Movie Database[3], we created a collaboration network of the movie issued in past decade (2000-2009), where each node represents a person who took part in the realization of a movie (directors, cast, song writers, and so on), and two persons are connected if they participated to the realization of the same movie. We considered each year as a dimension of the network.

Basic statistics of these networks are reported in Table I.

C. Quantitative Evaluation

Purpose of this section is to give a quantitative analysis of the results obtained, under two different perspectives driven by the following questions:

Q1. Can we evaluate the performances of the different conjunctions of ϕ and CD, and compare them among the different networks?

Q2. How does the choice of a combination of ϕ and CD affect the distribution of ρ over the communities?

In order to answer Q1, we looked at the values of the modularity measure (as defined in [6]), computed on the resulting set of communities C. Note that we could have computed the modularity on $\mathcal{C}$ instead (the modularity allows to be computed also in multidimensional networks), but this would have been inconsistent with the use of ϕ, which would have been disregarded in that way. Instead, the modularity takes into account the weights defined in ϕ.

This measure gives a value between zero and one, indicating how "good" nodes where partitioned into groups. The higher

[1] http://igraph.sourceforge.net/

[2] http://www.start.umd.edu/gtd

[3] http://www.imdb.com/

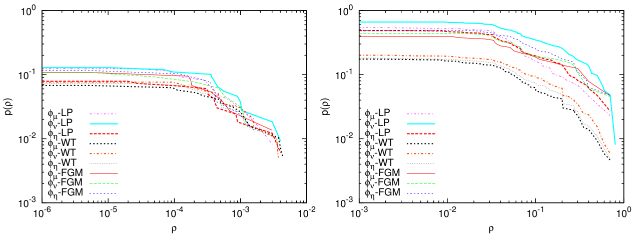

Fig. 2. The cumulative distribution for ρ in GTDand IMDb (Color Image).

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. Many researchers use the modularity scores as evaluation, or as parameter to be optimized by the community discovery algorithm. However, this is only a partial evaluation of the results, since the well-known problems of modularity [8] (such as the resolution problem, witnessed also by our Table II in which one can see that modularity-based algorithm FGM retrieve always a smaller number of bigger communities).

We computed anyway the modularity scores of each combination of community discovery algorithm and preprocessor, for all datasets. The results are reported in Table II, for every combination of network, ϕ, and CD. In the table we report in bold, for each algorithm, which ϕ produced the highest value of modularity. We are interested in seeing whether a specific combination of ϕ and CD tends to produce higher scores. Note that the values are not comparable between different networks since different network topologies may facilitate higher scores.

From Table II, we note that in only one out of six network-algorithm combinations, ϕ_μ was the best among the three ϕ. This confirms that, in most cases, keeping more information about the dimensions of a network leads to higher modularity, i.e. to a better set of communities.

In order to answer Q2, we analyzed the distribution of ρ for the output of each network-ϕ-algorithm combination. These distribution is depicted in Figure 2. We can see that 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 ρ measure.

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 ρ characterize the multidimensional structure of the partitioning.

VI. Conclusions and Future Work

We have addressed the problem of community discovery, applied to the scenario of multidimensional networks. We have given a possible definition of multidimensional community and provided a measure aimed at characterize the communities found. On this basis, we have devised a framework for finding and characterizing multidimensional communities, which is based on a mapping from multidimensional to monodimensional network, on the application of existing monodimensional community discovery algorithms to it, on the restoring of the originally residing multidimensional structure of the communities, and on the characterization of them via the ρ measures. Our results obtained on real world networks are encouraging, and provide a basis for future research on this direction. In particular, we plan to investigate the possibility of creating a multidimensional community discovery algorithm driven by ρ scores, possibly based on existing multidimensional methods such as the one in [12].

Acknowledgments. Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship. We thank Filippo Volpini for helping us with the R code.

References

[1] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multi-scale complexity in networks. *Nature*, 2010.

[2] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio A. F. Almeida. Characterizing user behavior in online social networks. In *Internet Measurement Conference*, pages 49–62, 2009.

[3] Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Mining the temporal dimension of the information propagation. In *IDA*, pages 237–248, 2009.

[4] Bringmann Bjoern, Berlingerio Michele, Bonchi Francesco, and Gionis Arisitdes. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4):26–35, 2010.

[5] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J.STAT.MECH.*, page P10008, 2008.

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

[7] Diane J. Cook, Aaron S. Crandall, Geetika Singla, and Brian Thomas. Detection of social interaction in smart spaces. *Cybernetics and Systems*, 41(2):90–104, 2010.

[8] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Science*, 104:36–41, January 2007.

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

[10] Steve Gregory. Finding overlapping communities using disjoint community detection algorithms. In *Complex Networks: CompleNet 2009*, pages 47–61. Springer-Verlag, May 2009.

[11] Jure Leskovec, Kevin J. Lang, and Michael W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW*, pages 631–640, 2010.

[12] 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.

[13] David L. Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM '03*, pages 556–559, New York, NY, USA, 2003. ACM.

[14] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[15] Spiros Papadimitriou, Jimeng Sun, Christos Faloutsos, and Philip S. Yu. Hierarchical, parameter-free community discovery. In *ECML PKDD '08*, pages 170–187, Berlin, Heidelberg, 2008. Springer-Verlag.

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

[17] 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.

[18] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105:1118–1123, January 2008.

[19] Lei Tang and Huan Liu. Scalable learning of collective behavior based on sparse social dimensions. In *CIKM*, 2009.

[20] Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. ICDM '02, pages 721–. IEEE Computer Society, 2002.