First publ. in: Advances in Data Analysis and Classification 5 (2011), 2, pp. 81-94

Network ensemble clustering using latent roles

Ulrik Brandes · Jürgen Lerner · Uwe Nagel

Abstract We present a clustering method for collections of graphs based on the assumptions that graphs in the same cluster have a similar role structure and that the respective roles can be founded on implicit vertex types. Given a network ensemble (a collection of attributed graphs with some substantive commonality), we start by partitioning the set of all vertices based on attribute similarity. Projection of each graph onto the resulting vertex types yields feature vectors of equal dimensionality, irrespective of the original graph sizes. These feature vectors are then subjected to standard clustering methods. This approach is motivated by social network concepts, and we demonstrate its utility on an ensemble of personal networks of migrants, where we extract structurally similar groups and show their resemblance to predicted acculturation strategies.

Keywords Social network analysis · Clustering · Network ensembles · Acculturation

Mathematics Subject Classification (2000) 62H30 · 91D30

1 Introduction

Clustering and classification of graphs have wide-spread applications in diverse fields such as pattern recognition, drug discovery, or biometrics. Clustering of graphs is of

U. Brandes · J. Lerner · U. Nagel (✉)
Department of Computer and Information Science, University of Konstanz, Konstanz, Germany
e-mail: Uwe.Nagel@uni-konstanz.de

U. Brandes
e-mail: Ulrik.Brandes@uni-konstanz.de

J. Lerner
e-mail: lerner@inf.uni-konstanz.de

course very different from graph clustering, sometimes also referred to as community detection, because it is not about decomposing a single graph (Schaeffer 2007; Fortunato 2010), but about detecting groups of similar graphs in a set of graphs. Our specific interest in and approach to this problem is motivated by the following scenario.

In the social sciences, subjects are often divided into social categories based on numerical or categorical personal attributes such as age, gender, race, job position, or income. Additional meaningful information, however, is given by an individual's personal network, i.e., his or her social contacts and the relations among them. In social network analysis, associations between attributes and social structure are of particular interest. For instance, it has been shown that the structure of personal networks correlates with psychological indicators (Kalish and Robins 2006). Likewise, personal networks have been used to define user roles in Usenet newsgroups (Welser et al. 2007) or to characterize the acculturation of migrants (Brandes et al. 2008; Molina et al. 2008). Note that in these applications one has to analyze and compare a *set* of networks rather than a single instance (see, e.g., Faust and Skvoretz 2002; Butts and Carley 2005; Faust 2006). Following Brandes et al. (2009), we refer to a set of networks that originate from the same underlying process, such as sampling or repeated measurement, as a *network ensemble*.

Clearly, networks in a given ensemble could be compared or characterized based on any one network characteristic or graph invariant such as density, clustering coefficient, or degree distribution. In this paper we compare attributed graphs using normalized projections onto vertex types, which simplify and standardize individual networks by exploiting an assumed relationship between vertex attributes and structural positions. Our comparisons are thus based on two aspects simultaneously: (1) "who is in the network", i.e., which individual characteristics do neighbors in the network have and (2) "how are they connected", i.e., what is the overall structure of ties among neighbors.

The remainder of this paper is organized as follows. Our method is described in detail in Sect. 2, followed by an illustrative application in Sect. 3. The decisions and assumptions leading to our method, as well as connections with random graph models and other clustering approaches are discussed in Sect. 4. We conclude with a brief outlook in Sect. 5.

2 Method

Assume we are given a collection of attributed graphs $\mathcal{G} = \{G_1 = (V_1, E_1), \ldots, G_N = (V_N, E_N)\}$ where $V_i = \{v_1^i, \ldots, v_{n_i}^i\}$ is the set of vertices of the ith graph and $E_i \subseteq \binom{V_i}{2}$, i.e., edges are unordered pairs of distinct vertices so that the graphs are simple and undirected. Vertex sets need not be disjoint. For convenience we assume that each vertex $v \in \mathcal{V} = \bigcup_{i=1}^{N} V_i$ is characterized by a real-valued attribute vector $\mathbf{a}(v) \in \mathbb{R}^d$ and we are given an attribute dissimilarity $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Our approach generalizes, however, to arbitrary graphs and attribute spaces.

The fundamental assumption underlying our method is that graphs from the same class are characterized by similar patterns of connectivity among similar types of vertices, and that there is variation in these patterns between classes.

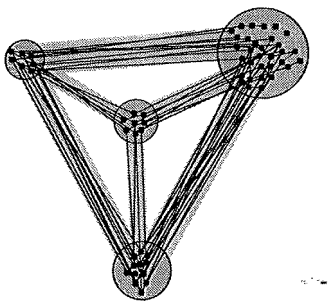

Fig. 1 Role structure resulting from the projection of a *graph* to four given types. Types are depicted as large *circles*, whereas vertices of the original *graph* are depicted as *black dots* and placed inside their type

Vertex types are represented as a partition $\mathcal{T} = \{T_1, \ldots, T_M\}$ of the overall vertex set $\mathcal{V}$, i.e., $\bigcup_k T_k = \bigcup_i V_i$ and $T_k \cap T_l = \emptyset$ if $k \neq l$. By contracting the vertices of each $G_i \in \mathcal{G}$ into at most M nodes representing their type, we obtain N simplified structures defined on a common vertex set $\mathcal{T}$. These so-called role structures are described by feature vectors with a joint signature and can therefore be clustered using common approaches. This yields a clustering of the ensemble $\mathcal{G}$ into groups of structurally similar graphs. As a byproduct, groups can be represented by their summarizing role structure.

We begin with a description of how a given partition of the overall vertex set can be used to embed the graphs in a feature space in Sect. 2.1. Clustering of an ensemble is described in Sect. 2.3 and only in Sect. 2.4 we address the partition of vertices defining their types.

2.1 Projections

Given a partition of the overall vertex set $\mathcal{V}$ into types $\mathcal{T}$, a *projection* maps each graph $G_i = (V_i, E_i)$ of an ensemble $\mathcal{G}$ to the complete (multi)graph defined on $\mathcal{T}$. Every vertex $v \in V_i$ is mapped to its corresponding type $T \in \mathcal{T}$ with $v \in T$, and every edge $e = \{v, w\} \in E_i$ to the edge $\{T, T'\}$ with $v \in T$ and $w \in T'$. For both vertices and edges of the image graph, multiplicities are counted.

While any partition of vertices into types induces such a projection, our central assumption is that the partition is such that—across an entire class of networks—vertices of the same type are connected are connected to other vertices in similar ways with respect to their neighbors' types. In other words, vertices of the same type are expected to assume the same *role* in the graph. The image of a projection is therefore referred to as a *role structure*. An example is shown in Fig. 1, where a graph is mapped to a role structure on four types using the projection that is induced by a partition into these types. Note that our use of the term 'role structure' is motivated more strongly by the theoretical concepts in Nadel (1957) than by the formal notions reviewed in Lerner (2005).

2.2 Features

To derive a fixed-length feature vector from a given projection of each graph $G_i \in \mathcal{G}$ we suggest to use properties such as the number of vertices of each type, the degree of

connectedness between and within types, and the total number of vertices in G_i. Properties thus fall into three categories: they are characteristics of types, of connections between types, or of the entire graph G_i. In each category, the number of attributes obtained is independent of the graph size, and fully determined by the role structure. That is, given a vertex partition, our feature vectors embed all graphs in a common feature space.

While the exemplary features discussed below have some plausibility, we note that others may be more useful in a given application context.

2.2.1 Cardinalities

The distribution of types among vertices in G_i is an important structural aspect for graph comparison, especially since types are induced by vertex attributes and therefore provide a specific substantive interpretation. To turn the distributions of vertex types into a feature, we add the relative frequency of occurrence of each type in the vertex set of graph as a component to the feature vector. Other size-related features that may be of interest are the numbers of vertices and edges in each G_i.

2.2.2 Connectivity

A second group of features reflects the importance of the distribution of edges between the vertices of specific types in G_i. For normalization purposes we do not simply count the number of edges between and within types, but we normalize them with the geometric mean of the involved type set cardinalities. Note that for equal-sized classes the geometric mean is proportional to the number of edges. In the case typical for social networks the average degree is constant and the number of edges thus linear in the number of vertices. For a graph $G_i = (V_i, E_i)$ and vertex types T_r and T_s this feature is defined as

$$e_{r,s} = \frac{|\{\{u, v\} \in E_i : u \in T_r \text{ and } v \in T_s\}|}{\sqrt{|V_i \cap T_r| \cdot |V_i \cap T_s|}},$$

i.e., we count the number of edges between vertices of types T_r and T_s in this graph and reweight them in such a way that the ratio of edge weights scales with average degrees. This scaling behavior is considered advantageous over that of standard density, and we will refer to it in the following as *average degree*. The upper triangle $(e_{r,s})_{1 \le r \le s \le M}$, including the diagonal ($M = |\mathcal{T}|$), of scaled multiplicities yields $\frac{1}{2}M(M+1)$ additional components of the feature vector encoding parts of the role structures of each graph. Comparability is ensured by fixing the ordering of vertex types.

The two feature groups sketched above are meant as examples only, since many others are conceivable. These feature vectors provide an embedding of the ensemble into a common space, consisting of subspaces for each group of property. Moreover, feature vectors can be utilized as signatures of graphs, and, because of their compatibility, prototypical signatures representing subsets of graphs can be derived as well. This is illustrated in our example application in Sect. 3.

One caveat is in place, however. When using the raw feature vectors as constructed in the last section, we face the problem that groups of features form subspaces of vastly different dimensionality and extent. A Euclidean distance on the combined space is therefore prone to be dominated by one or several of the subspaces. It may also be desirable to use individual distances on the subspaces or to emphasize the influence of certain features in a subsequent clustering of the ensemble.

Network ensemble clustering using feature vectors from role structures is therefore no different from other approaches based on vectors with inhomogeneous components. In a generic approach to moderate the effects of inhomogeneity, we propose to start by allowing arbitrary distances for each subspace. Further, distances are normalized such that the there is an expected unit-distance between two networks in every subspace, and finally we introduce weights when combining the subspace distances. More formally, let $\mathbf{f}(i)$ be the feature vector of graph G_i, $1 \leq i \leq N$, and let P be the set of feature groups $p \in P$. Then let $\mathbf{f}_p(i)$ denote the components of the feature vector corresponding to features in p. The normalized, weighted distance between two graphs (V_i, E_i) and (V_j, E_j) is then defined as follows:

$$\Delta(G_i, G_j) = \sum_{p \in P} \frac{\alpha_p}{\langle \delta_p \rangle} \cdot \delta_p(\mathbf{f}_p(i), \mathbf{f}_p(j))$$

where α_p is a weight for this feature group, δ_p the distance used in the subspace defined by p, and $\langle \delta_p \rangle$ is the average distance in this subspace over all pairs of graphs.

2.3 Clustering

Given feature vectors for the graphs in an ensemble, the identification of groups of structurally similar graphs can be solved as a standard clustering problem on this set of vectors.

In previous work we showed that graphs drawn from different planted partition models are separated well by their spectra (Brandes et al. 2009). Note that planted partition models essentially consist of an expected role structure. In the present situation, the existence of variation in role structures is only an assumption grounded on theoretical considerations, but since roles are actually defined by vertex types, we can make use of the role structures directly rather than indirectly via its showing in the spectrum.

For the specific features discussed in Sect. 2.2 it is expected indeed that in an ensemble of well-separable graphs we can find cluster representatives such that cardinalities and edge multiplicities of cluster members match well with their representatives, but display differences with other role structures.

While the actual choice of clustering method ultimately depends on hypothesized role structures and features chosen accordingly, one general requirement is implied by the underlying assumption of representative role structures. The selected clustering method should have a tendency towards compact, spherical clusters, for otherwise there is no feature vector that represents the core trends of all cluster members sufficiently

well. Since we are interested in relating categorical traits of persons and corresponding personal networks, there is no need for hierarchical clustering.

Clustering approaches such as k-means or the estimation of mixtures of Gaussian distributions as described in Fraley and Raftery (2002) generally seem appropriate in this scenario. Even if clusters are not well-separated, such methods yield reference points relative to which the individual personal networks can be interpreted.

2.4 Vertex types

A crucial building block of our approach is the plausible assignment of vertices to types, reflecting the way vertices connect to others in different classes of graphs (which are to be discovered in the ensemble).

For empirical studies, however, it seems reasonable to expect that the observed attribute data sufficiently discriminate the vertices (for otherwise we would be working with the implicit assumption that individuals can be distinguished from their relations alone), and at the same time display enough regularity to allow for the assignment of more general types (for otherwise there would be no association between personal attributes and relations).

Our approach is therefore to cluster all vertices based on their attributes. Since the attributes comprise all that we know about the entities represented by vertices, except for their relations, two vertices are indistinguishable, if they have identical attributes, and moderately different behavior is expected if attributes are similar. Because of this assumption, clusters of vertices should have small maximum distance between any pair of cluster members. In metric spaces, this implies that all members are near a common center of the cluster.

Since these requirements are similar to those discussed above, an estimation process for Gaussian mixture models as described in Fraley and Raftery (2002) may be appropriate. In general, of course, any suitable clustering method could be used for the determination of types as long as it yields compact clusters in the above sense.

3 An empirical real-case example

In the following we will describe the data set that will be the object of an example analysis. The example analysis starts with a vertex partition based on expert knowledge that will be described after the description of the data set. In Sect. 3.2 we show an analysis of this ensemble using our method and discuss the results in the context of acculturation strategies. A more detailed description of this specific application was previously given in Brandes et al. (2010).

3.1 Data

The collected data consists of a number of personal networks collected from migrants in Spain and the USA with the help of the EgoNet[1] software. Each of the 504 networks

[1] see http://www.egoredes.net for a description of the project and http://www.mdlogix.com/egonet.htm for a description of the software.

describes the social surrounding of a migrant to Spain or the USA, originating from a South-American, Middle-American, African or East-European country. Since the underlying data is equal to the data set used in Brandes et al. (2008), we reproduce that data description. Each respondent was asked to provide the following four types of information:

1. (*questions about ego*) 70 questions about the respondent herself, including age, skin color, years of residence, questions from traditional acculturation scales and health related questions.
2. (*name generator*) A list of 45 persons (referred to as *alters*) personally known to the respondent. The alters are represented as vertices in the respondent's personal network.
3. (*questions about alters*) 12 questions about each of the 45 alters, including country of origin, country of residence, skin color, and type of relation to ego.
4. (*ties between alters*) For each of the 990 undirected pairs of alters, the evaluation whether they know each other. The three possible choices were "very likely", "maybe" or "unlikely" and we introduced an edge in the network only if the respondent chose "very likely".

From a statistical point of view this seems to be a very small data set for the individual respondents. For our method it offers the opportunity to show the applicability on this kind of data. Through the combined analysis of the complete ensemble of networks we are able to identify classes of structurally similar networks thereby yielding a compact, abstracted description of the whole ensemble.

In the present example the only attribute data we will use are the countries of origin and residence of the alters and the same information about ego. By combination, vertex attributes describe the immigration situation of each alter (the vertices in each personal network) relative to the ego defining the network as in Brandes et al. (2008). hence, the alters of all networks are partitioned into four types:

- *origin* the alter stems from the the same country as the ego and still lives there.
- *fellows* the alter stems from and immigrated to the same country as the ego.
- *host* the alter lives in and stems from the country the ego immigrated to.
- *transnationals* all other.

In the following we will use these assignments of the vertices to types for the derivation of a structural description of the ensemble. When compared to the results of Brandes et al. (2008) for individual networks, the outcome of this process is a simple, unifying structural description of all networks in the ensemble.

3.2 Analysis

The first step of our analysis yields a description of all networks in the ensemble at once. This structural summary is visualized in Fig. 2. In addition to the descriptive values of cardinalities and average degree, the figure gives a visual impression of the average network structure. Edges are shaded according to the average degree described by them, that is the darker an edge is the higher is the average degree between vertices of the these types. Correspondingly, vertex intensity encodes the average degree

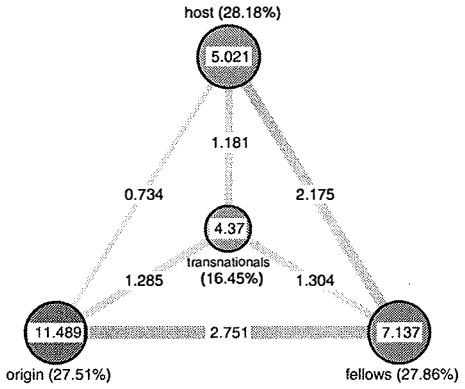

Fig. 2 Aggregate role structure in the ensemble of personal networks. Vertex sizes correspond to type frequency, where the share of vertices of each type is given in parenthesis. Intensity of edges and vertices corresponds to density inter- and intra-type connectivity

among vertices of the same type, while the average number of vertices of certain type is represented in the size of the node.

This description gives an overview of the ensemble by averaging over all networks. Additional measures such as standard deviation or descriptions of outliers could be added. Even from this simple representation, some general trends can be read that appear to hold throughout the ensemble. The most obvious detail is that the individual positions do not seem to differ much in size, except for the category of "transnationals". As expected the average degree within a type exceeds that between different types of vertices and the links "origin"-"fellows" and "host"-"fellows" are much stronger than those between other types. This can obviously be explained by straightforward arguments based on geographic distribution and homophily, but the diagram provides supporting evidence and quantifications.

In the following we will divide the ensemble into groups of structural similar networks and use this visualization method to give an overview of the characteristic features of each part. Together, these visualizations provide insight into the structure of the whole ensemble.

3.2.1 Role structure of individual clusters

Following the described method, we derived feature vectors for the networks, clustered them, and determined the role structure for each cluster. Visualizations of these role structures are shown in Fig. 3. These finer descriptions illustrate differences between clusters of networks and thereby trends in the ensemble determined by the clustering of the previous step. The feature vectors were constructed using the cardinality-related features introduced in Sect. 2.2. For clustering we used the k-means approach with k varying between 2 and 20, each repeated 1,000 times with different random initializations. The final partition was determined using the silhouette coefficient.

3.2.2 Interpretation of structural trends

As a result of this first examination we can associate the four clusters, corresponding to the four structural descriptions in Fig. 3, with the modes of acculturation proposed by Berry (1997). The networks belonging to Cluster 1 (Fig. 3a) show strong separation,

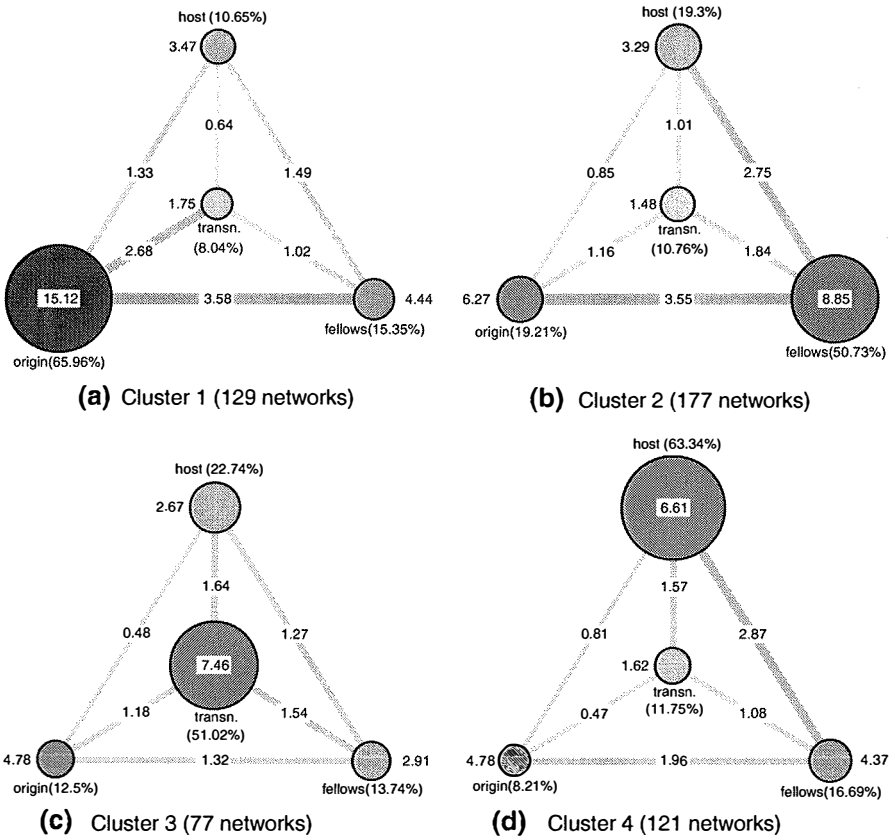

Fig. 3 Aggregate role structures of different clusters in the ensemble. Internal density of vertices of the same type is shown next to the type node, average degrees between types are given as labels on the edges. The fraction of actors belonging to a position is given in *parenthesis*

both with respect to nationality (most of their alters are born in the country of origin) and with respect to place of residence (most of their alters still live in the country of origin). The migrants giving rise to the networks in the second cluster (Fig. 3b) know many people living in the host country but still show strong separation with respect to nationality since most of their contacts are classified as fellow migrants. The persons in Cluster 3 can be interpreted to be well integrated; while the considerable number of "hosts" (about 22%) shows a good connection to the new society, none of the distinguished groups but the "transnationals"-type dominates these networks, so at least no concentration on a certain culture is observable. Migrants classified into Cluster 4 are assimilated since they know only few alters from their country of origin but most alters stem from the host society.

4 Discussion

In addition to other approaches for clustering sets of graphs, our work is also related to probabilistic network models that relate structure with vertex characteristics.

We discuss this line of research first, because it provides the main rationale of our approach.

4.1 Probabilistic network models

The assumption that vertex attributes determine edge probabilities is closely connected to the social space model of Boguñá et al. (2004). Here, an underlying (social) space is assumed in which the individuals corresponding to the vertices of the graph can be placed using their attributes. The probability of an edge between two vertices is then assumed to be directly related to their proximity in the social space; the smaller the distance, the higher is the probability of an edge. Thereby properties such as homophily and transitivity are inherently enforced. This is a feature that may be desired, but may also be limiting.

By summarizing similar vertex positions in types, we achieve that connection probabilities depend directly on the positions, not only on the distance of two vertices. The social distance approaches can thereby be connected with block models as in Holland et al. (1983) or planted partition models as defined in McSherry (2001). In these block models vertices are also assigned to types — the blocks — and the probability of an edge between two vertices depends only on the block memberships of the vertices. Holland et al. (1983) give some properties of block models and show how the parameters of such a model can be estimated for a graph with a priori known blocks.

A lossless transformation can be achieved by using a block for each distinct position in social space and derive the connection probabilities from the corresponding distances. By such a transformation every social space model can be expressed as a block model. Additionally, block models permit probabilities that are not embeddable in a metric space, thereby allowing to drop the assumptions of homophily and transitivity of edges. Finally, the summarization of vertex positions by types reduces noise in the data and establishes a natural matching of vertices between different graphs that can be used as the basis for a comparison.

4.2 Unsupervised learning

Similar graph clustering problems are encountered in areas such as pattern recognition, image analysis, or drug discovery. In addition to research on the unsupervised learning problem we faced here, much attention has been devoted to supervised learning or classification problems. In the following we will describe several approaches for the unsupervised learning problem which also produce cluster descriptions and additionally give some references to approaches concentrating on distances or similarities between graphs.

An algorithm to find correspondences from a set of graphs to an image graph is given in Heil and White (1976). The image graph consists of blocks and requirements on the connections between blocks. The aim is to find simultaneous homomorphisms mapping the vertices of the graphs to the image graph, while preserving the structural requirements of the image. Though no classification of an ensemble is achieved, this method is strongly related to ours both in its derivation from questions of social

sciences and through the assumption that the coarse-grained structure of a graph can be described by groups of similarly connected vertices.

4.2.1 Clustering based on probabilistic graph models

A specific notion of an ensemble is that of a sample from a mixture of probabilistic graph models. Wong and You (1985) and Wong et al. (1990) propose a probabilistic graph model together with a synthesis process that creates such a model to describe a group of attributed graphs. Their model consists of an underlying model graph with probability distributions for the attributes of its vertices and edges, where a special null-attribute encodes the absence of an element. To simplify the fit of such a model to an ensembles it is assumed that vertex and edge distributions are independent from each other and only the distribution of edges depends on attributes of the adjacent vertices (first order random graphs). They define a distance between attributed graphs, models and groups of attributed graphs based on entropy changes in the distributions described by the corresponding models. Determination of this distance involves finding of an optimal vertex mapping between two graphs. In another line of research it is argued that these models are over generalizing and alternative simplifications on dependencies between the individual distributions are introduced. Function described graphs are defined that have probabilities for vertex and edge attributes which are completely independent of each other and additional dependencies for realizations are introduced. Serratosa (2000) provides a good introduction and a number of references describing this approach.

4.2.2 Graph kernels: similarities between graphs

A very popular approach often used in supervised learning are *kernel methods*. In this framework it is sufficient to provide a similarity for the objects under examination to apply a number of algorithms of which the support vector machines seem to be the most popular. Consequently, a number of such similarities (kernels) have been defined to compare graphs with each other. Gärtner (2003) provides a survey on some of the different approaches and some hardness results are established in Gärtner et al. (2003). Here, additionally a kernel based on random walks is defined that uses walks with equal label sequences in both graphs for comparison. The random-walk kernel is based on the product graph, which pairs up vertices of both graphs using attributes as matching information. Neuhaus and Bunke (2006) propose an improvement of the random-walk kernel by pruning the product graph with a vertex matching obtained from the calculation of an edit distance. In Horváth et al. (2004) it is shown that also label sequences on cycles in the graphs can be used for comparison. Since the number of cycles can grow quite fast, this approach is only applicable for a certain class of graphs which is further extended in Horváth (2005). The usage of arbitrary frequent subgraphs is examined in Deshpande et al. (2005). Another possibility to define a similarity on graphs is given in Jain and Wysotzki (2004) and Jain et al. (2005). Here a product on the adjacency matrices is defined, that is the minimal Schur-Hadamard product under all permutations. The evaluation of this product involves an optimal vertex matching which is solved by a Hopfield network.

4.2.3 Distances

Within the different distances on graphs the edit distance as defined by Bunke and Allermann (1983) is one of the most popular ideas. Unfortunately the calculation of this distance is based on an optimal mapping between the vertices of the involved graphs.

Since methods for graph comparison generally suffer from this vertex permutation problem, the use of graph spectra is tempting. Consequently, in Luo et al. (2002, 2003) the authors describe experiments on the discriminatory qualities of a number of features based on graph spectra. They show that the leading eigenvalues have probably the best capabilities for structural comparison. Brandes et al. (2009) gives a possible explanation for this empirical result by proving that eigenvalues can be used to distinguish graphs on the different planted partition models they emerged from. It is shown that underlying planted partition models have an influence on the structure of the eigenvalues of randomly drawn graphs. Based on this, a distance is defined that distinguishes graphs by their underlying planted partition model. Unfortunately, the advantage that no attributes or vertex mappings are needed is accompanied by the calculation costs for the spectral decompositions and problems posed by small graphs and unclean matchings to the underlying models.

4.2.4 Abstraction using centers

Medians for sets of graphs are used for the abstraction and summarization of groups as in our method, but also in a number of clustering algorithms. To make algorithms based on centers such as k-means applicable to graphs, Bunke et al. (2003) constructs a super graph as cluster representation, and suggest a nearest neighbor clustering where graphs are added to clusters such that the change in entropy within the cluster is minimized. Defining the median of a set by an object that has the minimum sum of distances to all objects in the set a number of approaches exist that find such a median. Examples are Münger et al. (1999), Jiang et al. (1999, 2001), Hlaoui and Wang (2003, 2006). An alternative avoiding the need for medians is shown in Luo et al. (2001) where a clustering is found using an embedding of the graphs via multidimensional scaling based on arbitrary graph distances.

5 Conclusion and future work

We presented a method to cluster an ensemble of attributed graphs according to similarity in role structure, and compared it to existing approaches. At the core of our method is the definition of feature vectors based on the assumption that graphs differ by their role structure, and it can be used with any clustering algorithm respecting this assumption. Our approach naturally results in abstractions of individual networks and summarizing visualizations of role structures.

In its current form, our method requires the existence of a meaningful global partition of all vertices into types. It will be interesting to investigate ways to relax or even test this assumption without changing the random graph model. Assume, for

instance, an ensemble such as the one presented in Sect. 3 and a classification of its member networks that can be recovered without vertex attributes, but from connectivity information alone. Then, in turn, each network class model implies a partition of the vertices and can be used to investigate the relation between vertex attributes and structural positions, and possibly help identify attributes that align well with structural features.

Acknowledgments This research was supported in part by Deutsche Forschungsgemeinschaft under grants Br 2158/3-2 and GK 1042 (Research Training Group "Explorative Analysis and Visualization of Large Information Spaces"), the University of Konstanz under grant FP 626/08, and the European Commission through FP7-ICT-2007-C FET-Open Project BISON-211898. Method development was initiated and has benefitted substantially from cooperation with José Luis Molina, Miranda J. Lubbers, and Christopher McCarty, who also kindly let us use their data. We thank the anonymous reviewers and the editors for their helpful comments.

References

Berry JW (1997) Immigration, acculturation, and adaptation. Appl Psychol 46(1):5–68

Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. Phys Rev E 70(056122)

Brandes U, Lerner J, Lubbers MJ, McCarty C, Molina JL (2008) Visual statistics for collections of clustered graphs. In: Proceedings of the IEEE pacific visualization symposium (PacificVis'08). IEEE Computer Society, pp 47–54

Brandes U, Lerner J, Nagel U, Nick B (2009) Structural trends in network ensembles. In: Complex networks, volume 207 of studies in computational intelligence. Springer, pp 83–97

Brandes U, Lerner J, Lubbers MJ, McCarty C, Molina JL, Nagel U (2010) Recognizing modes of acculturation in personal networks of migrants. Procedia Soc Behav Sci 4:4–13 (Applications of Social Network Analysis)

Bunke H, Allermann G (1983) Inexact graph matching for structural pattern recognition. Pattern Recogn Lett 1(4):245–253

Bunke H, Foggia P, Guidobaldi C, Vento M (2003) Graph clustering using the weighted minimum common supergraph. In: Graph based representations in pattern recognition, volume 2726 of LNCS. Springer, pp 235–246

Butts CT, Carley KM (2005) Some simple algorithms for structural comparison. Comput Math Organ Theory 11(4):291–305

Deshpande M, Kuramochi M, Wale N, Karypis G (2005) Frequent substructure-based approaches for classifying chemical compounds. IEEE Trans Knowl Data Eng 17(8):1036–1050

Faust K (2006) Comparing social networks: size, density, and local structure. Metodološki zvezki 3(2): 185–216

Faust K, Skvoretz J (2002) Comparing networks across space and time, size and species. Soc Methodol 32(1):267–299

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97(458):611–631

Gärtner T (2003) A survey of kernels for structured data. ACM SIGKDD Explor Newsl 5(1):49–58

Gärtner T, Flach P, Wrobel S (2003) On graph kernels: hardness results and efficient alternatives. In: Proceedings of the 16th annual conference on computational learning theory and 7th kernel workshop, volume 2777 of LNCS. Springer, pp 29–143

Heil GH, White HC (1976) An algorithm for finding simultaneous homomorphic correspondences between graphs and their image graphs. Behav Sci 21(1):26–35

Hlaoui A, Wang S (2003) A new median graph algorithm. In: Graph based representations in pattern recognition, volume 2726 of LNCS. Springer, pp 225–234

Hlaoui A, Wang S (2006) Median graph computation for graph clustering. Soft Comput Fusion Found Methodol Appl 10(1):47–53

Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5(2):109–137
Horváth T (2005) Cyclic pattern kernels revisited. In: Advances in knowledge discovery and data mining, volume 3518 of LNAI. Springer, pp 791–801
Horváth T, Gärtner T, Wrobel S (2004) Cyclic pattern kernels for predictive graph mining. In: KDD '04: proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 158–167
Jain BJ, Wysotzki F (2004) Central clustering of attributed graphs. Mach Learn 56(1–3):169–207
Jain BJ, Geibel P, Wysotzki F (2005) SVM learning with the Schur-Hadamard inner product for graphs. Neurocomputing 64:93–105
Jiang X, Münger A, Bunke H (1999) Computing the generalized median of a set of graphs. In: Proceedings of the 2nd IAPR workshop on graph-based representations, pp 115–124
Jiang X, Münger A, Bunke H (2001) On median graphs: properties, algorithms, and applications. IEEE Trans Pattern Anal Mach Intell 23(10):1144–1151
Kalish Y, Robins G (2006) Psychological predispositions and network structure: the relationship between individual predispositions, structural holes and network closure. Soc Netw 28(1):56–84
Lerner J (2005) Role assignments. In: Brandes U, Erlebach T (eds) Network analysis. Springer, Berlin, pp 216–252
Luo B, Robles-Kelly A, Torsello A, Wilson RC, Hancock ER (2001) A probabilistic framework for graph clustering. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'01), vol 1. IEEE computer society, pp 912–919
Luo B, Wilson RC, Hancock ER (2002) Spectral feature vectors for graph clustering. In: Structural, syntactic, and statistical pattern recognition, volume 2396 of LNCS. Springer, pp 423–454
Luo B, Wilson RC, Hancock ER (2003) Spectral feature vectors for graph clustering. In: Graph based representations in pattern recognition, volume 2726 of LNCS. Springer, pp 190–201
McSherry F (2001) Spectral partitioning of random graphs. In: Proceedings of the 42nd annual IEEE symposium on foundations of computer science (FOCS'01). IEEE Computer Society, pp 529–537
Molina JL, Lerner J, Mestres SG (2008) Patrones de cambio de las redes personales de inmigrantes en Cataluña. Redes 15:50–63
Münger A, Bunke H, Jiang X (1999) Combinatorial search vs. genetic algorithms: a case study based on the generalized median graph problem. Pattern Recogn Lett 20(11–13):1271–1279
Nadel SF (1957) The theory of social structure. Cohen & West. Reprinted by Routledge, 2004
Neuhaus M, Bunke H (2006) A random walk kernel derived from graph edit distance. In: Structural, syntactic, and statistical pattern recognition, volume 4109 of LNCS. Springer, pp 191–199
Schaeffer SE (2007) Graph clustering. Comput Sci Rev 1(1):27–64
Serratosa F (2000) Function-described graphs for structural pattern recognition. PhD thesis, Institut d'Organització i Control de Sistemes Industrials
Welser HT, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. J Soc Struct 8
Wong A, You M (1985) Entropy and distance of random graphs with application to structural pattern recognition. IEEE Trans Pattern Anal Mach Intell 7:599–609
Wong A, Constant J, You M (1990) Random graphs. In: Bunke H, Sanfeliu A (eds) Syntactic and structural pattern recognition: theory and applications. World Scientific, pp 197–234