World Wide Web
DOI 10.1007/s11280-012-0190-4

Multidimensional networks: foundations of structural analysis

Michele Berlingerio · Michele Coscia · Fosca Giannotti · Anna Monreale · Dino Pedreschi

Received: 15 July 2011 / Revised: 6 February 2012 /
Accepted: 10 September 2012

Abstract Complex networks have been receiving increasing attention by the scientific community, thanks also to the increasing availability of real-world network data. So far, network analysis has focused on the characterization and measurement of local and global properties of graphs, such as diameter, degree distribution, centrality, and so on. In the last years, the multidimensional nature of many real world networks has been pointed out, i.e. many networks containing multiple connections between any pair of nodes have been analyzed. Despite the importance of analyzing this kind of networks was recognized by previous works, a complete framework for multidimensional network analysis is still missing. Such a framework would enable

This work was done when Michele Berlingerio and Michele Coscia were with KDDLab, ISTI - CNR, Pisa, Italy

M. Berlingerio (✉)
IBM Research, IBM Technology Campus, Damastown Industrial Estate, Dublin 15, Ireland
e-mail: mberling@ie.ibm.com, michele.berlingerio@isti.cnr.it

M. Coscia
Center for International Development, Harvard University, 79 JFK St, 02138 Cambridge, MA, USA
e-mail: coscia@di.unipi.it

F. Giannotti
KDDLab, ISTI - CNR, via G. Moruzzi 1, 56124 Pisa, Italy
e-mail: fosca.giannotti@isti.cnr.it

A. Monreale · D. Pedreschi
KDDLab, Dept. of Computer Science, University of Pisa,
largo B. Pontecorvo 3, 56100 Pisa, Italy

A. Monreale
e-mail: annam@di.unipi.it

D. Pedreschi
e-mail: pedre@di.unipi.it

Published online: 03 October 2012

the analysts to study different phenomena, that can be either the generalization to the multidimensional setting of what happens in monodimensional networks, or a new class of phenomena induced by the additional degree of complexity that multidimensionality provides in real networks. The aim of this paper is then to give the basis for multidimensional network analysis: we present a solid repertoire of basic concepts and analytical measures, which take into account the general structure of multidimensional networks. We tested our framework on different real world multidimensional networks, showing the validity and the meaningfulness of the measures introduced, that are able to extract important and non-random information about complex phenomena in such networks.

Keywords complex networks · social network analysis · World Wide Web

1 Introduction

In recent years, complex networks have been receiving increasing attention by the scientific community, also due to the availability of massive network data from diverse domains, and the outbreak of novel analytical paradigms, which pose at the center of the investigation relations and links among entities. Examples are social networks [3, 8, 14, 16], technology networks [2, 12], the World Wide Web [21, 28], biological networks [24, 25], and so on. 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, 31], detecting and studying the diffusion of information among them [23, 39], mining frequent patterns of users' behaviors [7, 20, 38, 40], are only a few examples of problems studied in Complex Network Analysis, that includes, among all, physicians, mathematicians, computer scientists, sociologists, economists and biologists.

Most of the networks studied so far are monodimensional: there can be only one link between two nodes. In this context, network analytics has focused on the characterization and measurement of local and global properties of such graphs, such as diameter, degree distribution, centrality, connectivity—up to more sophisticated discoveries based on graph mining, aimed at finding frequent subgraph patterns and analyzing the temporal evolution of a network.

However, in the real world, networks are often multidimesional, i.e there might be multiple connections between any pair of nodes. Therefore, multidimensional analysis is needed to distinguish among different kinds of interactions, or equivalently to look at interactions from different perspectives. This is analog to multidimensional analysis in OLAP systems and data warehouses, where data are aggregated along various dimensions. In analogy, we refer to different interactions between two entities as *dimensions*.

Dimensions in network data can be either *explicit* or *implicit*. In the first case the dimensions directly reflect the various interactions in reality; in the second case, the dimensions are defined by the analyst to reflect different interesting qualities of the interactions, that can be inferred from the available data. This is exactly the distinction studied in [29], where the authors deal with the problem of community discovery. In their paper, our conception of multidimensional network is referred as *multislice*, networks with explicit dimensions are named *multiplex*, and the temporal information is used to derive dimensions for the network.

Examples of networks with explicit dimensions are social networks where interactions represent information diffusion: email exchange, instant messaging services and so on. An example of network with implicit dimensions is an on-line social network with several features: in Flickr, while the social dimension is explicit, two users may be connected implicitly by the sets of their favorite photos.

Moreover, different dimensions may reflect different types of relationship, or different values of the same relationship. This is exactly the distinction reported in Figure 1, where on the left we have different types of links, while on the right we have different values (conferences) for one relationship (for example, co-authorship).

To the best of our knowledge, however, the literature still misses a systematic definition of a model for multidimensional networks, together with a comprehensive set of meaningful measures, that are capable of characterizing both global and local analytical properties and the hidden relationships among different dimensions. This is precisely the aim of this paper: we develop a solid repertoire of basic concepts and analytical measures, which take into account the general structure of multidimensional networks, with the aim of answering questions like: what is the degree of a node considering only a given set of dimensions? How are two or more dimensions related to each other? What is the "redundancy" among all the dimensions? To what extent one or more dimensions are more important than others for the connectivity of a node?

Our contribution can be then summarized as follows:

- we introduce a few examples of real-world multidimensional networks;
- we formally define a set of measures aimed at extracting useful knowledge on multidimensional networks;
- we empirically test the meainingfulness and scalability of our measures, by means of an extensive case study on the networks presented.

Our analysis shows that the measures we define are both simple and meaningful, and open the way for a new chapter of complex network analysis.

We extend our previous work [10] by adding more measures to the framework, increasing our set of networks to embrace a wider range of real world scenarios, and including a study on real world application scenarios in which we show the meaningfulness of our measures.

The rest of the paper is organized as follows: in Section 2 we present a few examples of real-world multidimensional networks; Section 3 introduces the measures

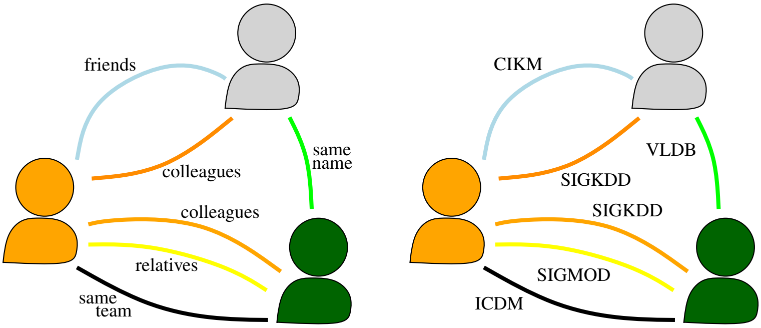


Figure 1 Example of multidimensional networks.

we define in this paper; Section 4 reports the empirical resuts obtained during our case study on real-world networks; in Section 5 we review a few related works; we conlude the paper in Section 6.

2 Multidimensional networks in reality

In the world as we know it we can see a large number of interactions and connections among information sources, events, people, or items, giving birth to complex networks. Enumerating all the possible networks detectable within our world, or their properties, would be difficult due to their number and heterogeneity, and it is not the scope of this paper. An excellent survey on complex network can be found in [30], where the author gives a good classification of networks into *social* (where, for example, we find on-line social network such as Facebook), *information* (such as for example citation networks), *technological* (among which we mention the power grid, the train routes, or the Internet), and *biological* (e.g., protein interaction networks) networks.

While all the example networks presented in [30] are monodimensional, in the real world it is possible to find many multidimensional networks. A few possible examples are:

Transportation Network. If we think about the complete transportation network of a country (or the world), we can easily see that we can build a multidimensional network where nodes represent the cities, and each transportation mean is a dimension. In this way, each city is connected to all the other cities reachable from it by means of airplanes, or buses, or trains, or ferries, or any kind of other available mean. As one can imagine, there will possibly be pairs of cities connected by more than one mean (e.g. Paris is connected to Madrid by both train and airplane), cities connected to the rest of the network by many means, or just one of them (think about cities on islands). It is interesting to note that we are, in turns, used to "browse" this network in its multidimensionality each time we travel: we take a train or a bus to reach the airport, then we flight from a city to another one, then we take another transportation mean to reach our final destination. It is clear also how this network is an aggregation of monodimensional networks corresponding to any single transportation mean, and that, according to our classification given in the previous section, this is a network in which dimensions reflect different types of explicit connections.

Social Network. Most of us nowadays use on-line services such as Facebook,[1] Flickr,[2] Skype,[3] Google+,[4] and so on. It is very common to have an account on many of them, because they provide different features, or we find different friends on them, or for any other reason. Each of us has a different user id in each of the networks, but if we join all the ids for every user, we can easily build a multidimensional social network, where any pair of people are connected by their friendship within

[1] http://www.facebook.com

[2] http://www.flickr.com

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

[4] http://plus.google.com

the different monodimensional networks. Significantly, there exist several multi-platforms to connect a single user to his/her multiple accounts at the same time (Pidgin,[5] Fring,[6] or Nimbuzz[7] are a few examples). As for above, two nodes here are connected by different types of connections, but in this case the links are not necessarily explicit. Two users for example may be linked in Flickr just because they use the same set of tags, or they like the same pictures, even if they are not explicitly friends.

Co-authorship Network. The aim of every conference is to gather together researcher in one particular area or topic. If we connect two authors by the papers they write together, it is clear to see that each conference, taken as dimension, provides its edges among the authors. There are, however, authors that publish on the same set of conferences for most of their collaborators, while others (mostly senior researchers) whose interests span multiple fields or topics, leading then to having a different set of neighboring collaborators for different dimensions. In this case, given the type of connection be the co-authorship, different conferences are different values of the links connecting the authors.

Utility Network. Most of our houses are connected to each other, or to main nodes, via different utility networks: water pipes, electric cables, phone and tv cables, build in fact a multidimensional network in which we live every day, where each utility is a technological network connecting different houses and offices. While at the node level this multidimensional network is highly redundant (almost every node is served by every utility), the network structure (i.e., the distribution of the links) might differ. In addition, this network also presents meta-nodes and hyperedges, due to the presence of pipe or cable junctions, network routers, utility headquarters, and so on.

The above is only a short, non-exhaustive list of possible real-world networks. Many other examples such as biological networks, other kinds of technological networks, social networks, peer-to-peer networks, and so on, can be found in reality.

2.1 Collected networks

While the above examples all are interesting and representative of a wide class of real-world networks with their properties, issues, and application scenarios, collecting data about them is not trivial and sometimes impossible. We present here a few multidimensional networks built from different dataset collected from various sources. The examples are real-world multidimensional networks, highly heterogeneous and representative of the possible different kinds of networks in the real world. We use these networks in the rest of the paper, to test our measures and to give possible application scenarios.

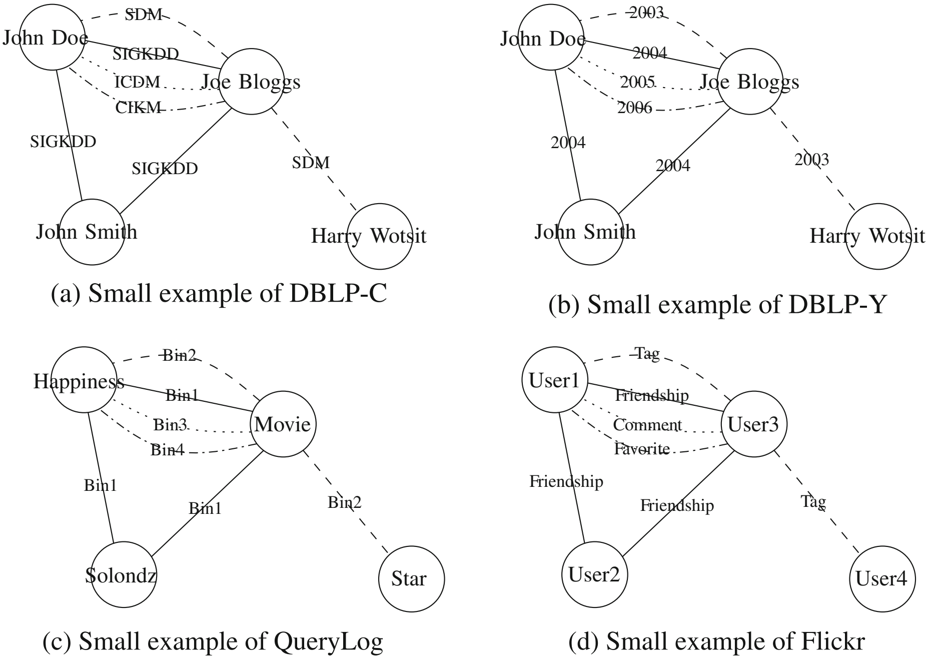
DBLP-C. We created this network from the well known bibliographic database DBLP.[8] We created a co-authorship network where the publication venues are

[5] http://www.piding.im

[6] http://www.fring.com

[7] http://www.nimbuzz.com

[8] http://www.informatik.uni-trier.de/~ley/db



(a) Small example of DBLP-C

(b) Small example of DBLP-Y

(c) Small example of QueryLog

(d) Small example of Flickr

Figure 2 Small extracts of the multidimensional networks built.

used as dimensions. In this network, we considered some of the most important conferences in Data Mining: SIGKDD, ICDM, SDM, VLDB, SIGMOD and CIKM. The authors were connected in a specific dimension if they wrote at least one paper together in the corresponding conference. A small extract of this network is represented in Figure 2a.

DBLP-Y. From the same DBLP source, we built also a co-authorship network of authors, using years from 1955 to 2009 as dimensions, and connecting two authors (nodes) in a specific dimension if they wrote at least one paper together in the corresponding year. A small extract of this network is represented in Figure 2b.

QueryLog. This network was constructed from a query log[9] of approximately 20 millions web-search queries submitted by 650,000 users, as described in [32]. Each record of this dataset stores a user ID, the query terms and the rank position of the result clicked by the user for the query. We extracted a word-word network of query terms (nodes), connecting two words if they appeared together in a query. The dimensions are defined as the rank positions of the clicked results, grouped into six almost equi-populated bins: “Bin1” for rank 1, “Bin2” for ranks 2–3, “Bin3” for ranks 4–6, “Bin4” for ranks 7–10, “Bin5” for ranks 11–58, “Bin6” for ranks 59–500. Hence two words appeared together in a query for which the user clicked on a resulting url ranked #4 produce a link in dimension “Bin3” between the two words. A small extract of this network is represented in Figure 2c.

[9] http://www.gregsadetsky.com/aol-data

Table 1 Basic statistics of the networks used: number of nodes, edges, dimensions, average degree, average number of neighbors.

Network	Dimension	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert D\rvert$	k	N
DBLP-C	VLDB	1,306	3,224		4.93	4.93
	SIGMOD	1,545	4,191		5.42	5.42
	CIKM	2,367	4,388		3.70	3.70
	SIGKDD	1,529	3,158		4.13	4.13
	ICDM	1,651	2,883		3.49	3.49
	SDM	915	1,501		3.28	3.28
	Total	6,771	19,345	6	5.71	5.04
DBLP-Y	Total	582,179	2,555,850	55	8.78	6.91
QueryLog	Bin1	138,991	1,104,581		15.89	15.89
	Bin2	108,438	878,136		16.19	16.19
	Bin3	89,417	708,897		15.85	15.85
	Bin4	75,845	583,774		15.39	15.39
	Bin5	42,950	253,976		11.83	11.83
	Bin6	12,235	36,456		5.96	5.96
	Total	184,760	3,565,820	6	38.60	19.26
Flickr	Friendship	984,919	48,723,010		98.93	98.93
	Comment	930,526	198,309,709		426.23	426.23
	Favorite	380,992	674,488,956		3,540.69	3,540.69
	Tag	91,690	715,447		15.60	15.60
	Global	1,186,895	922,237,122	4	1,554.03	1,455.62

Note that k and N are equivalent when computed on one single dimension

Flickr.[10] This dataset comes from the well known photo sharing service, and was obtained by crawling the data via the available APIs. We extracted both implicit and explicit dimensions of the social network represented in this data. For each picture, we extracted the list of all the users related to it and from these users we completed the social network by adding edges if two users commented, tagged or set the same picture as favorite, or if they had each other as a contact.

The resulting network is a person-person network, where each dimension is one of the "Friendship", "Tag", "Favorites", or "Comment", representing if the users are friends, tagged the same picture, marked the same picture as favorite, or commented the same picture. A small extract of this network is represented in Figure 2d.

Note that while for QueryLog we created our concept of dimensions, that are then to be considered implicit, in DBLP the authors explicitly set their collaborations, then the dimensions are explicit. In turns, Flickr has one explicit dimension (friendship) and three implicit (tag, favorites and comments). Moreover, in QueryLog, as well as in DBLP-Y, the dimensions reflect different quantitative values of the same type of relationships, while for DBLP-C and for Flickr the dimensions are built on different types of connections among users, and are not comparable.

Table 1 shows the basic properties of the networks, for each dimension, and for the total network. Note that k and N are equivalent when computed on a single

[10] http://www.flickr.com

dimension, and that DBLP-Y and DBLP-C have different aggregated values as they were built as different subsets of the entire DBLP data.

3 Multidimensional network analysis

In literature, many analytical measures, both at the local and at the global levels, have been defined in order to describe and analyze properties of standard, monodimensional networks. Defining meaningful measures provides several advantages in the analysis of complex networks. From the simplest measure, the *degree* of a node, to more sophisticated ones, like the *betweenness centrality*, or the *eigenvector centrality*, several important results have been obtained in analyzing complex networks on real-world case studies. These interesting network analytical measures come under a different light when seen in the multidimensional setting, since the analysis scenario gets even richer, thanks to the availability of different dimensions to take into account. As an example, the connectivity of the whole network changes if we see a single dimension as a separate network, with respect to the network formed by all the edges in the entire set of dimensions. Moreover, it would be interesting to analyze the importance of a dimension with respect to another, the importance of a dimension for a specific node, and so on. As a consequence, in this novel setting it becomes indispensable: (a) studying how most of the measures defined for classical monodimensional networks can be generalized in order to be applied to multidimensional networks; and (b) defining new measures, meaningful only in the multidimensional scenario, to capture hidden relationships among different dimensions.

Thus, in the remainder of this section, we introduce the elements composing our model as follows. First, we introduce a mathematical model for multidimensional networks. Although not being the only possibility (other possibilities would include tensors, among all), we found multigraphs to be a simple and versatile model, that allow also for a simple a fast implementation of the measures (see Section 4). Then, we discuss the extension of monodimensional measures to the multidimensional setting. For sake of simplicity, we only present one measure, the degree, although it is possible to extend most of the monodimensional measures following the same strategy of adding a parameter to the domain of the functions. Lastly, we introduce our multidimensional measures, meaningful only in the multidimensional setting. To give an overview, we introduce both measures that are local to the nodes, and measures that are global to the dimensions. The set of measures introduced is not meant to be complete: other measures can be defined, for example, at the intermediate level of the ego-networks, or they can be assessing links instead of nodes. For the sake of simplicity, however, we introduce only a few, generic, measures, together with toy examples meant to help understand their meaning, and we will explore in the future the possibility of introducing new ad-hoc measures that are meant to be used in specific application-driven contexts (for example, measures for evolving multidimensional networks, measures for semantic networks, and so on).

3.1 A model for multidimensional networks

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 $G = (V, E, L)$ where: V is a set of nodes; L 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 $d \in L$ 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 $d \in L$ only one edge (u, v, d) may exist. Thus, each pair of nodes in G can be connected by at most $|L|$ possible edges. Hereafter $\mathcal{P}(L)$ denotes the power set of L.

3.2 Extending monodimensional measures

How can we extend the analytical measures defined on monodimensional networks to deal with multiple dimensions? In general, in order to adapt the classical measures to the multidimensional setting we need to extend the domain of each function in order to specify the set of dimensions for which they are calculated. Intuitively, when a measure considers a specific set of dimensions, a filter is applied on the multigraph to produce a view of it considering only that specific set, and then the measure is calculated over this view. In the following, due to space constraints, we show how to redefine only the well-known *degree* measure by applying the above approach. Note that most of the classical measures can be extended in a similar way.

In order to cope with the multidimensional setting, we can define the degree of a node w.r.t a single dimension or a set of them. To this end, we have to redefine the domain of the classical degree function by including also the dimensions.

Definition 1 (Degree) Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network G, respectively. The function $Degree : V \times \mathcal{P}(L) \rightarrow \mathbb{N}$ defined as

$$Degree(v, D) = |\{(u, v, d) \in E \; s.t. \; u \in V \;\wedge\; d \in D\}|$$

computes the number of edges, labeled with one of the dimensions in D, between v and any other node u.

We can consider two particular cases: when $D = L$ we have the degree of the node v within the whole network, while when the set of dimensions D contains only one dimension d we have the degree of v in the dimension d, which is the classical degree of a node in a monodimensional network. This kind of consideration also holds for any measure that is possible to extend to the multidimensional case in this way.

In order to illustrate the measures we define in this paper, we use a toy example, depicted in Figure 3, to show the application of the measures on it.

Example 1 Consider the multigraph in Figure 3 that models a multidimensional network with 2 dimensions: dimension d_1 represented by a solid line, and dimension d_2 represented by the dashed line. In this multigraph we have $Degree(3, \{d_1\}) = 2$, $Degree(3, \{d_2\}) = 0$ and $Degree(2, \{d_1, d_2\}) = 3$.

3.3 Multidimensional measures

In this section we define new measures on the multidimensional setting and that are meaningful only in this scenario.

3.3.1 Neighbors

In classical graph theory the *degree* of a node refers to the connections of a node in a network: it is defined, in fact, as the number of edges adjacent to a node. In a simple graph, each edge is the sole connection to an adjacent node. In multidimensional networks the degree of a node and the number of nodes adjacent to it are no longer related, since there may be more than one edge between any two nodes. For instance, in Figure 3, the node 4 has five neighbors and degree equal to 7 (taking into account all the dimensions). In order to capture this difference, we define the following:

Definition 2 (Neighbors) Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively. The function $Neighbors : V \times \mathcal{P}(L) \rightarrow \mathbb{N}$ is defined as

$$Neighbors(v, D) = |NeighborSet(v, D)|$$

where

$$NeighborSet(v, D) = \{u \in V \mid \exists (u, v, d) \in E \ \wedge \ d \in D\}.$$

This function computes the number of all the nodes directly reachable from node v by edges labeled with dimensions belonging to D.

Note that, in the monodimensional case, the value of this measure corresponds to the degree. It is easy to see that $Neighbors(v, D) \leq Degree(v)$, but we can also easily say something about the ratio $\frac{Neighbors(v,D)}{Degree(v)}$. When the number of neighbors is small, but each one is connected by many edges to v, we have low values of this ratio, which means that the set of dimensions is somehow redundant w.r.t. the connectivity of that node. This is the case of node 5 in the toy example illustrated. On the opposite extreme, the two measures coincide, and this ratio is equal to 1, which means that each dimension is necessary (and not redundant) for the connectivity of that node: removing any dimension would disconnect (directly) that node from some of its neighbors. This is the case of node 2 in Figure 3.

We also define a variant of the Neighbors function, which takes into account only the adjacent nodes that are connected by edges belonging exclusively to a given set of dimensions.

Definition 3 (Neighbors$_{\text{XOR}}$) Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively. The function $Neighbors_{\text{XOR}}$:

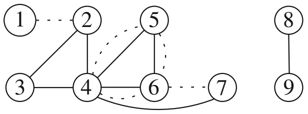


Figure 3 Toy example. *Solid line* is dimension 1, the *dashed* is dimension 2.

$V \times \mathcal{P}(L) \to \mathbb{N}$ is defined as

$$Neighbors_{\mathrm{XOR}}(v, D) = |\{u \in V|\ \exists d \in D : (u, v, d) \in E \wedge \nexists d' \notin D : (u, v, d') \in E\}|$$

It computes the number of neighboring nodes connected by edges belonging only to dimensions in D.

3.3.2 Dimension relevance

One key aspect of multidimensional network analysis is to understand how important a particular dimension is over the others for the connectivity of a node, i.e. what happens to the connectivity of the node if we remove that dimension. We then define the new concept of *Dimension Relevance*.

Definition 4 (Dimension relevance) Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively. The function $DR : V \times \mathcal{P}(L) \to [0, 1]$ is defined as

$$DR(v, D) = \frac{Neighbors(v, D)}{Neighbors(v, L)}$$

and computes the ratio between the neighbors of a node v connected by edges belonging to a specific set of dimensions in D and the total number of its neighbors.

Clearly, the set D might also contain only a single dimension d, for which the analyst might want to study the specific role within the network, to assess, for example, the importance of the single conference in DBLP-C over the others.

However, in a multidimensional setting, this measure may still not cover important information about the connectivity of a node. Figure 3 shows two nodes (4 and 5) with a high dimension relevance for the dimension represented by a solid line. Specifically, in both cases the dimension relevance is equal to one, but the complete set of connections they present is different: if we remove the dimension represented with a solid line, the node 4 will be completely disconnected from some its neighbors, for example it cannot reach the nodes 2, 3 and 7 anymore; while the node 5 can still reach all its neighbors. To capture these possible different cases we introduce a variant of this measure.

Definition 5 (Dimension relevance XOR) Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively. $DR_{\mathrm{XOR}} : V \times \mathcal{P}(L) \to [0, 1]$ is defined as

$$DR_{\mathrm{XOR}}(v, D) = \frac{Neighbors_{\mathrm{XOR}}(v, D)}{Neighbors(v, L)}$$

and computes the fraction of neighbors directly reachable from node v following edges belonging only to dimensions D.

Example 2 We can easily calculate the above measure for the nodes in Figure 3. As an example, for the node 8 there is no difference with the DR (Definition 4): all its neighbors are only reachable by solid edges. The opposite situation holds for node 5: all its neighbors are reachable by solid edges, but we always have an alternative edge. So the DR_{XOR} of the solid line dimension is equal to zero.

In the following, we want to capture the intuitive intermediate value, i.e. the number of neighbors reachable through a dimension, weighted by the number of alternative connections.

Definition 6 (Weighted dimension relevance) Let $v \in V$ and $d \in L$ be a node and a dimension of a network $G = (V, E, L)$, respectively. The function $DR_W : V \times \mathcal{P}(L) \to [0, 1]$, called *Weighted Dimension Relevance*, is defined as

$$DR_W(v, D) = \frac{\sum_{u \in NeighborSet(v, D)} \frac{n_{uvd}}{n_{uv}}}{Neighbors(v, L)}$$

where: n_{uvd} is the number of dimensions which label the edges between two nodes u and v and that belong to D; n_{uv} is the number of dimensions which label the edges between two nodes u and v.

Hereafter we occasionally use DRs to indicate all the three variants of this measure. Note that $DR_{\mathrm{XOR}} = 0$ does not necessary imply that the node is not connected to a particular dimension. It represents a situation where the node has no neighbors that can be reached exclusively through that particular dimension. So it is possible to reach it by alternative ways. In Figure 3, node 5 is an example of this, when considering the dashed (or solid) line dimension.

The Weighted Dimension Relevance takes into account both the situations modeled by the previous two definitions. Low values of DR_W for a set of dimensions D are typical of nodes that have a large number of alternative dimensions through which they can reach their neighbors. High values, on the other hand, mean that there are fewer alternatives. Our example shows the case of node 5 when considering the solid line dimension: its DR_W is clearly the highest, although the dashed line dimension has a high value of DR.

3.3.3 Highest and lowest redundancy connections nodes

We introduce two new concepts regarding the nodes of multidimensional networks: *Highest Redundancy Connections (HRC)* and *Lowest Redundancy Connection (LRC)* nodes. They are derived from the combination of the functions Degree and Neighbors. Intuitively, these measures describe the structure around a given node in terms of edge density: if the node is a LRC this structure is sparse, while if the node is HRC it is dense and redundant.

Definition 7 (LRC) A node $v \in V$ is said to be at *Lowest Redundancy Connection (LRC)* if each of its neighbors is reachable via only one dimension, i.e.,

$$\forall u \in NeighborSet(v, L) : \exists!\, d \in L\ (u, v, d) \in E.$$

Note that if a node v is LRC we have

$$Degree(v, L) = Neighbors(v, L).$$

Definition 8 (HRC) A node $v \in V$ is called *Highest Redundancy Connections (HRC)* if each of its neighbors is reachable via all the dimensions in the network, i.e.,

$$\forall u \in NeighborSet(v, L) : \forall d \in L\ (u, v, d) \in E.$$

Note that if a node v is HRC we have

$$Degree(v, L) = Neighbors(v, L) \times |L|.$$

Example 3 In Figure 3 we have several LRC nodes: 1, 2, 3, 7, 8 and 9. Some of them appear in both dimensions (2 and 7), while other nodes appear in only one dimension (1, 3, 8 and 9). On the other hand we have only one HRC node: node number 5 is connected via both the dimensions with each of its neighbors.

In the "utility network" introduced in Section 2, we have that most of the nodes are HRC, as most of the houses have electricity, water pipes, gas, and so on. On the other hand, in the "transportation network", little islands are most likely to be LRC, as most of them are connected to their neighboring cities only by ferry (excluding the ones with little airports).

3.3.4 Dimension connectivity

Another interesting quantitative property of multidimensional networks to study is the percentage of nodes or edges contained in a specific dimension or that belong *only* to that dimension. To this end we also introduce: the *Dimension Connectivity* and the *Exclusive Dimension Connectivity* on both the sets of nodes and edges.

Definition 9 (Node dimension connectivity) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $NDC : L \to [0, 1]$ defined as

$$NDC(d) = \frac{|\{u \in V \mid \exists v \in V : (u, v, d) \in E\}|}{|V|}$$

computes the ratio of nodes of the network that belong to the dimension d.

Definition 10 (Edge dimension connectivity) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $EDC : L \to [0, 1]$ defined as

$$EDC(d) = \frac{|\{(u, v, d) \in E | u, v \in V\}|}{|E|}$$

computes the ratio of edges of the network labeled with the dimension d.

Definition 11 (Node exclusive dimension connectivity) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $NEDC : L \to [0, 1]$ defined as

$$NEDC(d) = \frac{|\{u \in V \mid \exists v \in V : (u, v, d) \in E \land \forall j \in L, j \neq d : (u, v, j) \notin E\}|}{|\{u \in V \mid \exists v \in V : \ (u, v, d) \in E\}|}$$

computes the ratio of nodes belonging only to the dimension d.

Definition 12 (Edge exclusive dimension connectivity) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $EEDC : L \rightarrow [0, 1]$ defined as

$$EEDC(d) = \frac{|\{(u, v, d) \in E \mid u, v \in V \ \wedge \ \forall j \in L, j \neq d : \ (u, v, j) \notin E\}|}{|\{(u, v, d) \in E \mid u, v \in V\}|}$$

computes the ratio of edges between any pair of nodes u and v labeled with the dimension d such that there are no other edges between the same two nodes belonging to other dimensions $j \neq d$.

Example 4 In Figure 3 the EDC of dimension d_1 is 0.61 since it has eight edges out of the 13 total edges of the network. Its EEDC is equal to $5/8 = 0.625$. The NDC for the same dimension d_1 is 0.88 (8 nodes out of 9) and its NEDC is 0.375 (3 unique nodes out of 8).

Table 3 presents the values of these measures computed on our real-world networks.

3.3.5 D-Correlation

The last aspect of multidimensional networks that we study in this paper is the interplay among dimensions. In the following we define two measures that, intuitively, give an idea of how redundant are two dimensions for the existence of a node or an edge. These two measures are based on the classical Jaccard correlation coefficient, but they extend it in order to cope with more than two sets.

Definition 13 (Node D-Correlation) Let $D \subseteq L$ be a set dimensions of a network $G = (V, E, L)$. The *Node D-Correlation* is the function $\rho_{\text{nodes}} : \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$\rho_{nodes}(D) = \frac{|\bigcap_{d \in D} V_d|}{|\bigcup_{d \in D} V_d|}$$

where V_d denotes the set of nodes belonging to dimension d. It computes the ratio of nodes appearing in all the dimensions in D and the total number of nodes appearing in at least one dimension in D.

Definition 14 (Pair D-Correlation) Let $D \subseteq L$ be a set dimensions of a network $G = (V, E, L)$. The *Pair D-Correlation* is the function $\rho_{pairs} : \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$\rho_{pairs}(D) = \frac{|\bigcap_{d \in D} P_d|}{|\bigcup_{d \in D} P_d|}$$

where P_d denotes the set of pairs of nodes (u, v) connected in dimension d. It computes the ratio of pairs of nodes connected in all the dimensions in D and the total number of pairs of nodes connected in at least one dimension in D.

Figures 4, 5 and 6 show the behavior of these measures on our real-life networks.

When $D = L$, we can compute the percentage of nodes that exist in all the dimensions of the network, that we call *Omni-Connected Nodes (OCN)*, and, in analogy, the percentage of pairs of nodes connected in all the dimensions, that we call *Omni-Connected Pairs (OCP)*. Table 3 reports these percentages on our networks.

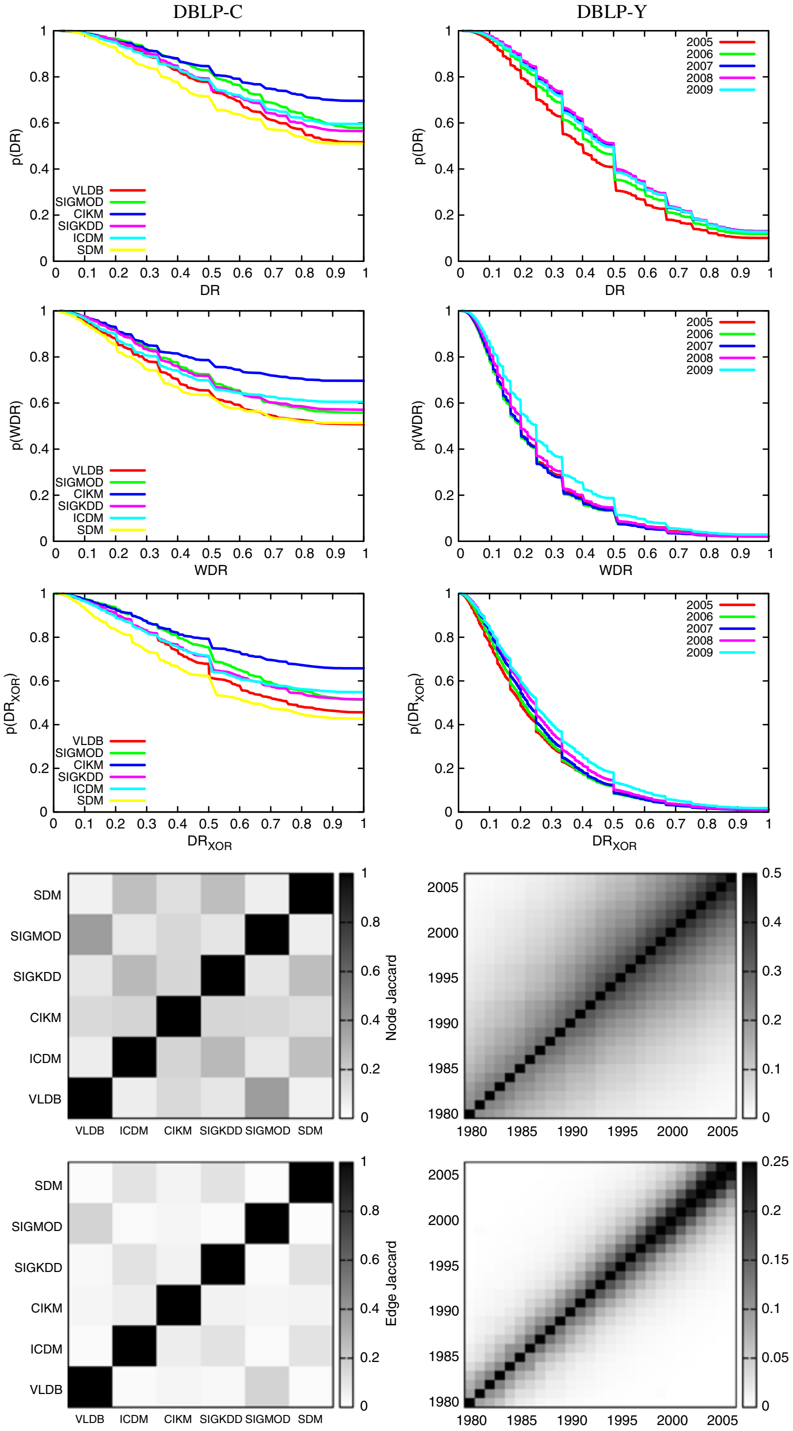


Figure 4 The cumulative distributions of the three DRs (*first three rows*), Node D-Correlation (*fourth row*) and Pair D-Correlation (*last row*) in DBLP-C and DBLP-Y.

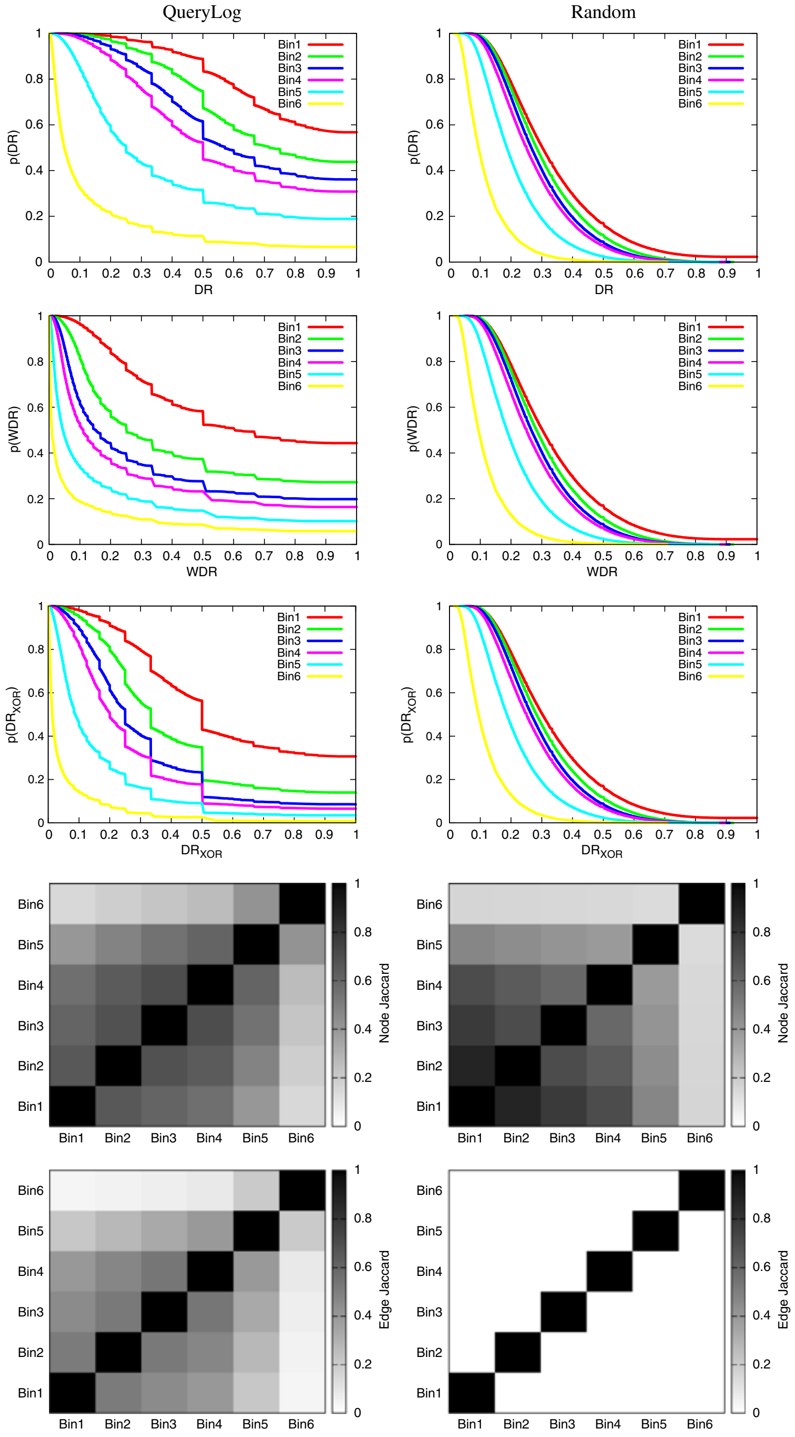
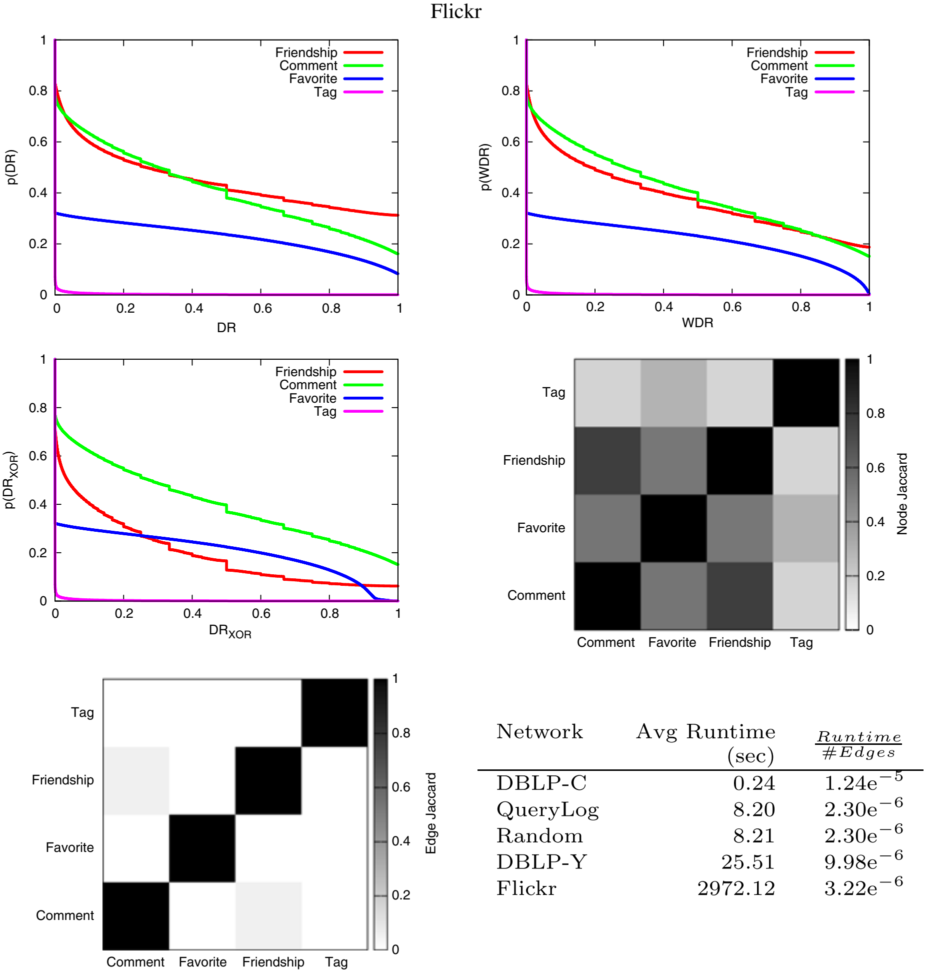


Figure 5 The cumulative distributions of the three DRs (*first three rows*), Node D-Correlation (*fourth row*) and Pair D-Correlation (*last row*) in QueryLog and Random.



Network	Avg Runtime (sec)	$\frac{Runtime}{\#Edges}$
DBLP-C	0.24	$1.24e^{-5}$
QueryLog	8.20	$2.30e^{-6}$
Random	8.21	$2.30e^{-6}$
DBLP-Y	25.51	$9.98e^{-6}$
Flickr	2972.12	$3.22e^{-6}$

Figure 6 First five figures: the cumulative distributions of the three DRs, Node D-Correlation and Pair D-Correlation in Flickr. Table in the bottom right: running times for computing the measures.

4 Experiments

In this section, we present the results obtained by computing all the defined measures on our real world networks presented in Section 2. In Table 2 we provide a summary of our measures and their abbreviations to make the reading of this section easier.

In order to better understand the meaning of our measures, we also created a random network to be used as null model for our experiments. The network was created at random, while preserving the basic characteristics (number of nodes and number of edges) of each single dimension of the QueryLog network. Thus, we call each of its dimensions with the name of the corresponding dimension in QueryLog, while we refer to the network as Random, or "null model".

All the experiments were conducted on a server equipped with a dual Xeon processor at 3.06 Ghz, 8 GB of RAM, and running Ubuntu 8.04 server 64 bit. We

Table 2 Summary of notation.

Notation	
DR	Dimension relevance
DR_{XOR}	Dimension relevance XOR
DR_W	Weighted dimension relevance
NDC	Node dimension connectivity
NEDC	Node exclusive dimension connectivity
EDC	Edge dimension connectivity
EEDC	Edge exclusive dimension connectivity
HRC	Highest redundancy connections
LRC	Lowest redundancy connections
OCP	Omni-connected pairs
OCN	Omni-connected nodes

discuss scalability in Section 4.1. An implementation in Java of all the measures presented is available for download.[11]

Figures 4–6 report, the cumulative distribution of the three variants of the DRs and the matrices of the Node D-Correlation and Pair D-Correlation for every pair of dimensions in the network (higher values mapped to darker color), respectively (since DBLP-Y has 55 dimensions, for clarity or space issue we report the values only for the last 5 dimensions). In Figures 4 and 5 each column contains plots about a specific network: the first three plots describe the different variants of the DR while the last two plots show the Node D-Correlation and Pair D-Correlation. Figure 6 contains all the plots about the Flickr network and a table about the information on the runtime for each network.

First thing to note about the DRs is that different networks present different distributions of these measures. In addition, it is easy to see that each network behaves differently from the null model. In particular, the distinction between the QueryLog and the Random network (Figure 5) is very clear, despite having used the statistics of QueryLog to build the null model. Different distributions are showing that the knowledge extracted on real networks is much different from the one extracted on a random one, i.e. we are not assessing a random phenomenon.

Now, we analyze the correlation between the DR distribution and the Dimension Connectivity values (especially the EEDC and NEDC). What can be seen by looking at the DR distributions, the EEDC and NEDC values, reported in Table 3, is that the DR distributions seem to be correlated to the EEDC measure while the DR_{XOR} distributions seem to be correlated to the NEDC. This correlation is not surprising since by definition, the two measures are two different perspectives, one local (Dimension Relevance) and one global (Dimension Connectivity), of the same aspect: how much a dimension is important for the connectivity of a network. We note, in fact, that the DR tends to be higher in conjunction with higher Edge Exclusive Dimension Connectivity values (e.g. in the DBLP-C network). This can be read as: distributions similar to those of the DBLP-C network (first column of Figure 4) occur when the dimensions are quite independent from each other. The QueryLog network (first column of Figure 5) presents much more separated distributions among the dimensions where the EEDC values present an high variance. Moreover,

[11] http://kdd.isti.cnr.it/MHA

Table 3 Dimension Connectivity, HRC, LRC, OCN and OCP computed on the used networks.

Network	Dim	NDC	NEDC	EDC	EEDC	HRC	LRC	OCN	OCP
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
DBLP-C	VLDB	19.28	0.75	16.67	74.75	0	79.58	0.18	0.01
	SIGMOD	22.81	0.97	21.66	80.02				
	CIKM	34.95	3.86	22.68	84.59				
	SIGKDD	22.58	1.38	16.33	78.68				
	ICDM	24.38	2.45	14.90	76.24				
	SDM	13.51	1.44	7.76	68.28				
DBLP-Y	2005	65.35	0.50	16.24	36.87	0.35	9.78	19.42	2.83
	2006	74.69	0.47	19.36	30.90				
	2007	78.81	0.47	21.34	29.78				
	2008	78.62	0.48	22.10	33.51				
	2009	75.01	0.58	20.96	42.33				
QueryLog	Bin1	75.22	12.58	30.98	38.47	0.04	42.47	3.14	0.78
	Bin2	58.69	4.39	24.63	22.39				
	Bin3	48.39	2.19	19.88	16.30				
	Bin4	41.05	1.41	16.37	14.05				
	Bin5	23.24	0.42	7.12	10.72				
	Bin6	6.62	0.02	1.02	4.45				
Random	Bin1	75.22	0	30.98	99.97	0	99.26	0.43	0
	Bin2	58.69	0	24.63	99.97				
	Bin3	48.39	0	19.88	99.96				
	Bin4	41.05	0	16.37	99.96				
	Bin5	23.24	0	7.12	99.96				
	Bin6	6.62	0	1.02	99.97				
Flickr	Friendship	82.98	71.07	5.28	2.07	2.94e^{-3}	29.54	5.75	8e^{-5}
	Comment	78.39	77.14	21.50	21.36				
	Favorite	32.09	32.08	73.13	63.89				
	Tag	7.72	7.51	0.08	0.07				

the descending order (by dimension) of EEDC follows the decreasing trend (by dimension) in the cumulative distribution plots.

However, the correlation between the EEDC and the DR is not evident in Flickr. This is manly due to the high unbalance among the dimensions: the "Favorite" dimension is clearly dominant in number of edges w.r.t the other ones. Instead, the numbers reveal that there is a correlation between the DR_{XOR} and the NEDC values: the trend of the NEDC values per dimension is followed by the DR_{XOR} distributions (see the third plot in Figure 6).

In Figures 4–6 we also report the values of the two correlations we defined. We recall that, due to the underlying Jaccard correlation, the matrices shown in these figures are symmetric. In these matrices, we reported the correlations computed on each possible pair of dimensions. The values computed on the complete set of dimensions, corresponding to the OCN e OCP percentages, are reported in Table 3.

In the last two rows of Figures 4 and 5 we see that the presence of a natural ordering among the dimensions lets a clear phenomenon emerge: closer dimensions are more similar than distant ones, according to the natural order. The phenomenon is highlighted by the fact that the cells close to the diagonal are darker than those distant from it, in Querylog and DBLP-Y networks. In these two networks, in fact, there is a natural order of the years and the bins, used as dimensions. This is not true for DBLP-C: it is not possible to establish a natural ordering among conferences, thus the corresponding matrices in the first column of Figure 4 appear to be more "random".

Consider now the matrices related to the Random network. Due to the random generation, the natural ordering of the dimensions disappears, while, in this case, the size of the dimensions does the difference, and the trend of the correlation follows the basic statistics of the network. Hence, more nodes and edges in a dimension imply more correlation with the other ones. This phenomenon is particularly easy to observe for the nodes: the number of possible edges is very large, thus it is difficult to create, using a random generator, the same edges in two different dimensions, dramatically lowering then the Pair D-Correlation values (which appears almost white in the second column of Figure 5) and bringing close to 100 % the EEDC values (NEDC values are all equal to zero due to the artifact of choosing the random node ids from the same set).

This is true also considering HRC and OCP values of our networks, reported in Table 3. The null model does not present any node with these properties, while instead it has the highest number of LRC nodes. This is again an effect of the above mentioned properties: too many edge combinations lead the edges of a random network to appear only in one dimension. On the other hand, in DBLP-Y we have some authors publishing each year with all their collaborators (HRC column) or at least one time each year (OCN column). These two events are quite rare in the random null model. Some networks may present also situations even more extreme than the random null model: it is the case of DBLP-C in which only 12 authors have published in all the six considered conferences (OCN column), and only two pairs have collaborated at least once in all the conferences (OCP column). But this is natural, since publishing in all of these top conferences is very difficult.

The matrices of Node D-Correlation and Pair D-Correlation of Flickr in Figure 6 are coherent with the number of nodes and edges in the different dimensions. Indeed, we can observe that the Pair D-Correlation values between the dimensions is very low as the number of edges per dimension in this network is very different (see Table 1). In contrast, for the Node D-Correlation values we have a different correlation matrix due to the fact that the number of nodes per dimension is very similar but for the dimension "Tag" (see Table 1). Indeed, we can see that for the dimension "Tag" we have the lowest correlation values. The particular interplay among the dimensions of this network and so the high difference in terms of number of edges in the different dimensions also affects the values of HRC and OCN that are very low. Indeed, it is hard to find an edge appearing in all the dimensions in Flickr.

Again, these considerations support the thesis that our multidimensional measures are capturing real, and not random, phenomena, that constitute meaningful knowledge mined in the multidimensional networks analyzed.

4.1 Scalability

Since all the measures introduced may be trivially computed by scanning only once the list of nodes and edges (theoretical discussion and implementation details omitted), the entire framework scales on both time and memory. For the latter, we report that the largest memory occupation was less than 600 MB. For the running times, the table in Figure 6 reports, for each network, the average running time (computed over ten executions) and the ratio between time and number of edges of the networks. While the time varies considerably among the networks, we see that the ratio with the number of edges is almost constant. One exception is the DBLP-C network, for which, however, the overhead payed by the initialization of the data structures is significant w.r.t the total running time (reported in seconds). This empirical evaluation shows that the measures can be efficiently computed, in line with the theoretical complexity which is linear in the number of edges.

4.2 Application scenarios

As we said above, we deal everyday with multidimensional networks, and thus we could enumerate an extensive list of application scenarios of multidimensional network analysis. For example, considering the transportation network described in Section 2, we note that when planning our movements, we implicitly solve an instance of the shortest path problem in the multidimensional transportation network, trying to minimize the cost paid in each dimension (both in terms of time and money), and the cost for changing transportation mean (i.e., the overhead given by the interplay of the dimensions).

Instead of reporting such a list, which would be impossible to fit in this work, and which is not its main purpose, we now give three examples of application scenarios on the networks used for our experiments, assessing the meaningfulness of our measures in the context of the World Wide Web and scientific publishing. The first deals with search engines and query terms, the second regards on-line social networks and finally we analyze publishing behavior of computer scientists.

Other more complex applications scenarios for the introduced model, namely multidimensional community discovery and characterization [9] and multidimensional link prediction [34], have been already investigated, and, in the future, we plan to address problems such as frequent subgraph mining, clustering, classification, similarity, and so on, driven by our introduced model.

4.2.1 Detection of ambiguous query terms

In the QueryLog network we apply our measures to find ambiguous query terms. In order to do so, we select the query terms that are: 1) used in conjunction with many other terms (high number of neighbors) and 2) generally connected with their neighbors in queries that led to low rank results (low Weighted Dimension Relevance for the first rank bin, i.e. the neighboring terms are often found in queries that do not provide good results for the user).

Then, we are saying that being an ambiguous query term translates into being a hub with a low value for DR_W, calculated on dimension “Bin1”. Note this choice: minimizing the DR_{XOR} of dimension “Bin1” would have selected terms that generally do not produce good results at all, while the pure DR would not have specified the interplay with the other dimensions.

First, we extract some nodes that are hubs (i.e., nodes with a high number of adjacent edges; please refer to [11] for a formal definition) in the network and then we consider the small communities of words surrounding these nodes extract where we look for the reasons for a very good or very bad query result. We select the neighbors with the highest Dimension Relevance for dimension 1, to see why, with a generally bad query term, sometimes we find good results.

A possible example found to satisfy these criteria is the word "Wearing" (a simplified view of its neighborhood is depicted in Figure 7a). This term shows here poor semantics, which needs a disambiguation. Moreover, the clusters surrounding this word are very clear: words in either cluster are not really expected to be in the other one. The first group of queries was apparently generated by users looking for information about AIDS and how to prevent it. In the second cluster we see people interested in Elle MacPherson's dressing habits. We found a total of 1150 nodes with a structure of neighbors similar to the one of "Wearing".

4.2.2 Outlier detection

Here we analyze a totally different context, i.e. a network of social connections. The aim of this analysis is to find users that are connected to the network mainly via the Friendship dimension, thus disregarding the Comment, Favorite and Tag features of the social network.

Thus, in this analysis we focused on the Dimension Relevance XOR and considered the head of its distribution for the Friendship dimension: higher values of this measure mean that the node is connected with its neighborhood exclusively via Friendship links.

Table 4 shows the values of the Dimension Relevance XOR of the Friendship for the 12 nodes obtained by maximizing both the DR_{XOR} for Friendship, and the number of Neighbors. The last two columns indicate the direction of the connection between each node and its neighbors (extracted a posteriori since our network was undirected).

We can identify two categories of users among the interesting extracted nodes: *professionals* and *spammers*, for which Figure 7(b) gives a possible representation. The first can be identified in the table due to their high number of ingoing edges and the low number of outgoing ones. This behavior is classic in social networks: if a person has an interesting profile, many people will ask for friendship (instances

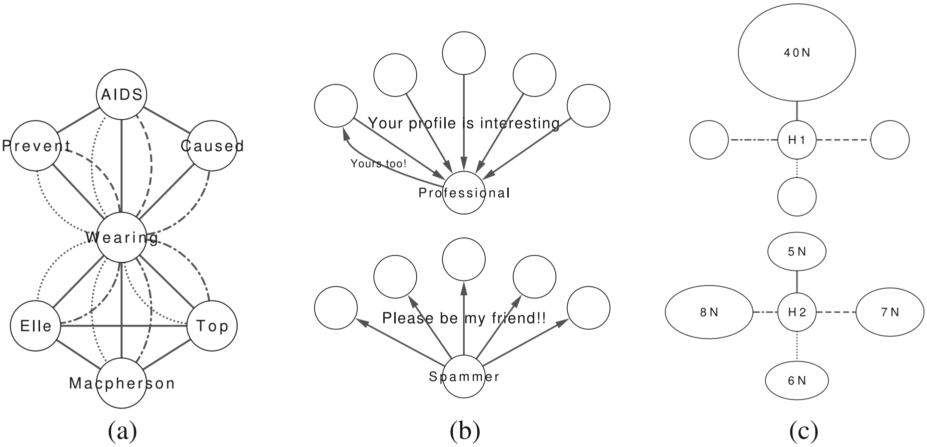


Figure 7 Some of the multidimensional hubs extracted.

Table 4 DR_{XOR} for 12 nodes extracted from Flickr maximizing neighbors.

Node	$\mathrm{DR}_{\mathrm{XOR}}$	Out	In
0	0.78	152	5080
1	0.79	3836	0
2	0.79	3766	0
3	0.79	8	8091
4	0.81	4203	0
5	0.82	3704	0
6	0.83	7226	0
7	0.83	655	4066
8	0.85	4205	0
9	0.86	750	6983
10	0.88	138	3671
11	0.95	4301	0

of this behavior are available in Flickr[12,13]). On the other hand, the owner of an interesting profile could not be interested in having so many friends (instances of this behavior are available in Flickr[14,15]). The opposite observation can be made for spammers: they can be detected by a high number of outgoing edges but no one is interested in returning the friendship link to a spammer. Note that looking only at the difference (or ratio) between outgoing and incoming links is not enough: we need the DR_{XOR} to filter the users for which the friendship has a relatively high relevance. This is clearly due to our definition of "spam", other definitions are also possible (for example, one can look at the number of personal messages sent, but this is out of scope for our purposes).

4.2.3 Analyzing temporal behaviors

In this section, we go beyond the theory presented so far. We want to compute the DRs considering not only one dimension at a time, but a set of many dimensions.

In this context, an interesting application of our approach is to analyze the temporal behavior of multidimensional hubs on evolving networks. In this section we show the results obtained on DBLP-Y, whose dimensions are the years of publications. Note that, although multi-dimensionality and temporal information can be both present in a network, in the case of monodimensional networks we can still use the temporal information to apply multidimensional techniques and highlight interesting phenomena.

The specific object of our analysis is to find authors of scientific papers who tend to change the authors with whom they collaborate possibly every year. Note that we are not focusing on just new collaborations, but we want also to see the old ones to disappear. In order to do so, we found hubs v maximizing the number of dimensions d for which $\mathrm{DR}_{\mathrm{XOR}}(v, d) > 0$ (maximizing this value means maximizing the number of years in which the author had collaborations that took place only in a specific year and not in others).

[12] http://www.flickr.com/photos/10539246@N05

[13] http://www.flickr.com/photos/23941584@N08

[14] http://www.flickr.com/photos/38687875@N00

[15] http://www.flickr.com/photos/20532904@N00

Figure 7c reports two representations of hubs extracted in this way: the hubs behaving as H1 and the ones behaving as H2. To be more precise, a deeper classification among them might be performed by looking also at the standard deviation of the DR_{XOR} computed in all the dimensions. The example H2 in the right of that Figure, in fact, represents a hub minimizing the standard deviation. H1 hubs are collaborators in high effort publications such as books (such as Maxine D. Brown or Steffen Schulze-Kremer); while H2 hubs are authors who tend work with many different people, rarely keeping these collaborations alive, such as Ming Xu or Jakob Nielsen.

5 Related work

In this section we briefly review some studies that are related to this paper under two different perspectives: first we review a few classical achievements of complex network analysis and then we go through possible models and measures for multidimensional networks.

An exhaustive survey of network analytics is provided by Newman [30], where it is shown how many properties apply to various kinds of networks that we find in the real world, spanning from social to biological networks; then the basic properties of networks are discussed: small world effect, clustering coefficient, degree distributions, network resilience, together with various network generation models. Network science is today a highly visible field of research, with relevant books also tailored for broad dissemination [5, 15, 37]. A large body of work was dedicated to the analysis of the degree distribution in networks, often with reference to specific networks such as phone calls [1], Internet [22], the Web [6, 27], citation networks [33], online social networks [18] and many others. One popular result is the power law distribution of the degree in many real-world networks. Another interesting survey paper by Chakrabarti and Faloutsos is [17], where, besides network properties, several graph generators are presented. The authors also give a review of basic concepts of graph mining (i.e., the problem of finding frequent subgraphs), navigation in graphs (crawling, search, and so on), generic flows in graphs (information, viruses, etc.), and possible applicative contexts of social networks in various fields, such as Viral Marketing (i.e. trying to individuate the smallest set of users that maximize the spread of advertisement) or Recommendation Systems.

Concerning multidimensional networks, there is little work so far on a general methodology for multidimensional network analysis, and a few works that address specific problems in a multidimensional setting.

The authors of [19] introduce the *graph OLAP*, a multidimensional view of graph data introduced with the purpose of defining the aggregation of different dimensions. However, a systematic definition of analytical measures is missing and the interplay among different dimensions is not investigated in any way. In other words, the graph OLAP is a method for supporting the navigation along the dimensions of a network, not a general framework for multidimensional network analysis.

Some recent works put emphasis on specific multidimensional social networks, such as, as an example, communication networks among people [36]. Given a network and a set of latent social dimensions the authors were able to determine how new entities will behave in these dimensions. In this paper, the authors focus on relational learning, extracting latent social dimensions via modularity maximization. Based on the extracted social features, a discriminative classifier like SVM is constructed to determine which dimensions are informative for classification. Although the underlying setting is similar to the one studied in our paper, the authors only focus on a particular problem, and develop specific analytical means for their objectives. Our attempt, in this paper, is precisely to find a suitable level of generalization that allows us to put into focus the truly important primitives for multidimensional network analysis, in order to devise a framework that can be systematically used in practice for addressing a wide variety of problems. Two more papers deal with the analysis of multidimensional network [28, 35]. In both cases, the authors analyze networks with "positive" and "negative" links among on-line communities. In [35], the authors analyze the degree distributions of the various dimensions, which are scale-free structures, highlighting the need for analytical tools for the multidimensional study of hubs. In [28], the authors presented the problem of predicting the positive (trust) or negative (distrust) label of the edges. While this might look like a multidimensional formulation of the LP problem, its formulation is, in turns, a classification problem, as only the label of the edge, rather than its future arrive, is predicted. In [26] the authors introduced a semi-supervised learning model for the link prediction problem in multi-relational networks. Like multidimensional networks, multi-relational ones allow different types of interactions between each pair of nodes. However, this model does not allow for multiple simultaneous interactions between two nodes.

Finally, in [29] the authors deal with the problem of community discovery, and extend the definition of modularity to fit to the multidimensional case, which they call "multislice".

A last set of related works deal with the problem of analyzing *heterogeneous* networks. In this class of networks, nodes can have different types, and multiple labels can be also associated with them. One work of this kind is [4], where the authors deal with the problem of graph-based classification in heterogeneous networks.

6 Conclusions and future work

In this paper, we studied the problem of analyzing *multidimensional* networks. We have introduced a large, solid, repertoire of meaningful measures, able to capture different interesting structural properties of multidimensional networks, such as the interplay residing among the dimensions, both at the global and at the local level. Aware that such an ambitious definitional apparatus needs to be empirically assessed, we devoted a large effort to gather multidimensional network data, and performed an extensive set of empirical experiments. We believe that the many experiments over massive, real-world network data from heterogeneous domains validated the sense and the analytical power of our repertoire of measures. According to our findings, our measures also appear to be able to capture real, non random, phenomena, and allow for interesting interpretation of the results.

On the other hand, we are aware that the research described in this paper leaves many problems open for further research, both on the theoretical and the application side. Is the repertoire of measures sufficiently wide to express the desired class of analytical questions? Are there interesting properties of the measures that may help the analysis, or be exploited for optimizing the computation of the measures themselves? What should be the characteristic of a query system capable of supporting the proposed analytical framework for multidimensional networks? These are the main challenges that we plan to pursue in the next future, along with continuing our field experiments over ever richer, larger and more complex network data.

Acknowledgements Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship.

References

1. Abello, J., Buchsbaum, A.L., Westbrook, J.R.: A functional approach to external graph algorithms. In: Algorithmica, pp. 332–343. Springer (1998)
2. Adamic, L.A., Lukose, R.M., Puniyani, A.R., Huberman, B.A.: Search in power-law networks. Phys. Rev. E **64**(46135) (2001)
3. Aiello, W., Chung, F., Lu, L.: A random graph model for massive graphs. In: STOC, pp. 171–180. ACM (2000)
4. Angelova, R., Kasneci, G., Weikum, G.: Graffiti: graph-based classification in heterogeneous networks. World Wide Web **15**, 139–170 (2012). doi:10.1007/s11280-011-0126-4
5. Barabási, A.L.: Linked: The New Science of Networks. Perseus Books Group (2002)
6. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509 (1999)
7. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.A.F.: Characterizing user behavior in online social networks. In: Internet Measurement Conference, pp. 49–62 (2009)
8. Berlingerio, M., Coscia, M., Giannotti, F.: Mining the temporal dimension of the information propagation. In: IDA, pp. 237–248 (2009)
9. Berlingerio, M., Coscia, M., Giannotti, F.: Finding and characterizing communities in multidimensional networks. In: ASONAM, pp. 490–494 (2011)
10. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D.: Foundations of multidimensional network analysis. In: ASONAM, pp. 485–489 (2011)
11. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D.: The pursuit of hubbiness: analysis of hubs in large multidimensional networks. J. Comput. Sci. **2**, 223–237 (2012)
12. Berlingerio, M., Pinelli, F., Nanni, M., Giannotti, F.: Temporal mining for interactive workflow data analysis. In: KDD pp. 109–118 (2009)
13. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A.: Learning and predicting the evolution of social networks. IEEE Intell. Syst. **25**, 26–35 (2010)
14. Musial, K., Kazienko, P.: Social networks on the internet. World Wide Web J. (2012). doi:10.1007/s11280-011-0155-z
15. Buchanan, M.: Nexus: Small Worlds and the Groundbreaking Theory of Networks. W.W. Norton & Co. (2003)
16. De Castro, R., Grossman, J.W.: Famous trails to Paul Erds. Math. Intell. **21**, 51–63 (1999)
17. Chakrabarti, D., Faloutsos, C.: Graph mining: laws, generators, and algorithms. ACM Comput. Surv. **38** (2006)
18. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-mat: a recursive model for graph mining. In: ICDM (2004)
19. Chen, C., Yan, X., Zhu, F., Han, J., Yu, P.S.: Graph olap: towards online analytical processing on graphs. In: ICDM, pp. 103–112 (2008)
20. Cook, D.J., Crandall, A.S., Singla, G., Thomas, B.: Detection of social interaction in smart spaces. Cybern. Syst. **41**(2), 90–104 (2010)
21. Donato, D.: Graph structures and algorithms for query-log analysis. In: Ferreira, F., Löwe, B., Mayordomo, E., Mendes Gomes, L. (eds.) CiE, Lecture Notes in Computer Science, vol. 6158, pp. 126–131. Springer (2010)
22. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: SIGCOMM, pp. 251–262 (1999)
23. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: KDD, pp. 1019–1028 (2010)

24. Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature **411**(6833), 41–42 (2001)
25. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
26. Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., Tsuda, K.: Link propagation: a fast semi-supervised learning algorithm for link prediction. In: SDM, pp. 1099–1110. SIAM (2009)
27. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The Web as a Graph: Measurements, Models, and Methods (1999)
28. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: WWW, pp. 641–650. ACM (2010)
29. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.-P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**, 876 (2010)
30. Newman, M.E.J.: The Structure and Function of Complex Networks (2003)
31. Nowell, D.L., Kleinberg, J.: The link prediction problem for social networks. In: CIKM ’03, pp. 556–559. ACM (2003)
32. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: InfoScale ’06, p. 1. ACM (2006)
33. Redner, S.: How popular is your paper? an empirical study of the citation distribution. Eur. Phys. J., B Cond. Matter Complex Syst. **4**(2), 131–134 (1998)
34. Rossetti, G., Berlingerio, M., Giannotti, F.: Scalable link prediction on multidimensional networks. In: Spiliopoulou, M., Wang, H., Cook, D.J., Pei, J., Wang, W., Zaïane, O.R., Wu, X. (eds.) ICDM Workshops, pp, 979–986. IEEE (2011)
35. Szell, M., Lambiotte, R., Thurner, S.: Trade, conflict and sentiments: multi-relational organization of large-scale social networks. PNAS 107(31), 13636–13641. arXiv.1003.5137 (2010)
36. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: KDD, pp. 817–826. ACM (2009)
37. Watts, D.J.: Six Degrees: The Science of a Connected Age (2003)
38. Yan, X., Han, J.: gspan: graph-based substructure pattern mining. ICDM ’02, pp. 721–724 (2002)
39. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) ICDM, pp. 599–608. IEEE Computer Society (2010)
40. Zhao, P., Yu., J.: Fast frequent free tree mining in graph databases. World Wide Web **11**, 71–92 (2008). doi:10.1007/s11280-007-0031-z