Noname manuscript No.
(will be inserted by the editor)

Foundations of Multidimensional Network Analysis

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

the date of receipt and acceptance should be inserted later

Abstract 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 relations and links among entities, or people, at the center of investigation. Networks are usually modeled by graphs. So far, network analytics has focused to the characterization and measurement of local and global properties of such graphs, such as diameter, degree distribution, centrality, connectedness - 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 practice, real networks come with a rich semantics attached to relations, and nodes in a network may be connected by edges of different nature: for example, any given pair of persons may communicate with different tools (phone, email, messaging, etc), or in a social network can be linked by a different relation (being friends, colleagues, relatives, etc). A network where several possible connections (edges) exist between the same pair of entities (nodes) is called a *multidimensional network*. Despite the importance of this kind of network is recognized in many works, and ad-hoc analytical means have been proposed to deal with multidimensional networks of specific cases, a thorough systematic framework for multidimensional network analysis is still missing. This is precisely the aim of this paper: we develop a solid repertoire of basic concepts and analytical mechanisms, which takes into account the general structure of multidimensional networks: first, we model a multidimensional network as a *multigraph*, i.e., a graph where nodes can be connected by one or more labeled edges; second, we systematically develop a vast repertoire of network metrics for the graph, to characterize local and global properties of multidimensional networks. We show how popular measures like the degree of a node, the number of connected components in a graph, the shortest path, and so on, can be viewed as particular cases of more general definitions for multidimensional networks. Further, we introduce brand new metrics for multigraphs, that take into consideration the interplay among different dimension, and therefore have no counterpart in the single-dimension case. In order to demonstrate the usefulness and wide applicability of the proposed framework, we consider a large array of massive networks in diverse domains, ranging from query logs to social networks, customer networks, subgraphs and bibliographic networks, and show how in each such case the introduced metrics - both the generalization of the known ones and the brand new multidimensional metrics - reveal a surprising high analytical power and suggest novel solutions to challenging real life problems.

M. Berlingerio, F. Giannotti
KddLab - ISTI CNR
Via G. Moruzzi, 1, 56124 Pisa - Italy
E-mail: {michele.berlingerio, fosca.giannotti}@isti.cnr.it

M. Coscia, A. Monreale, D. Pedreschi
KddLab - Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3, 56127 Pisa - Italy
E-mail: {coscia, annam, pedre}@di.unipi.it

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 relations and links among entities, or people, at the center of investigation. In fact,

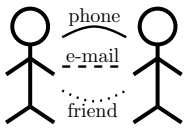


Fig. 1 Example of multidimensional network

in as different fields as mathematics, physics, computer science, biology, chemistry, economy and sociology, we can find networks of users, or items, or any kind of interacting entities, which are worth studying for many interesting purposes [**?**,**?**,**?**,**?**,**?**,**?**,**?**].

Networks are usually modeled by graphs. So far, network analytics has focused to the characterization and measurement of local and global properties of such graphs, such as diameter, degree distribution, centrality, connectedness - 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 practice, real networks come with a rich semantic residing in the relations, and nodes in a network may be connected by edges of different nature: for example, any given pair of persons may communicate with different tools (phone, email, messaging, etc), or in a social network can be linked by a different relation (being friends, colleagues, relatives, etc). A network where several possible connections (edges) exist between the same pair of entities (nodes) is called a *multidimensional network*. Figure **??** shows a possible multidimensional network, constituted by only two users, where three dimensions ("phone", "email", "friend") connect them. It is worth noting that a multidimensional network may contain heterogeneous dimensions, and this is the case in the figure, where communication tools and social relations coexist together.

The necessity of multidimensional network analysis is recognized by several authors. Previous work, for instance, introducted the *OLAP graph* paradigm, which provides multidimensional and multilevel views of graph data [**?**]. However, this work concentrates on defining means for drilling up and down some aggregated measures along different dimensions, while a comprehensive definition of the aggregate measures for multidimensional network analytics is missing, as well as the means to explore the relations among the different dimensions.

More recent works put emphasis on the analysis of specific multidimensional social networks such as, as an example, communication networks among people [**?**]. Here, the authors focus on a particular problem, and define adequate metrics to this purpose. This situation recurs in many studies which, explicitly or not, involve multidimensional networks: ad-hoc notions are put forward, to the end of characterizing some interesting aspect of the network under analysis.

What is missing, then, is a systematic definition of 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 and dependencies among different dimensions. This is precisely the aim of this paper: we develop a solid repertoire of basic concepts and analytical mechanisms, 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?
- What are the connected components of the graph considering two specific dimensions?
- How are two or more dimensions related to each other? To what extent does one of them "contain" the other?
- What is the "redundancy" among all dimensions?
- Are there dimensions that are good representative of the entire network, and that would allow us to have a condensed representation of the data?

The contributions of this paper can be hence summarized as follows. First, we model multidimensional networks with the *multigraph*, i.e., a graph where nodes are connected by one or more labeled edges.

Second, we review the network metrics known in the literature, such as the degree of a node, the number of connected components in a graph, the shortest path, and so on, and we show how they can be viewed as particular cases of more general definitions for multidimensional networks.

After this first step, we define new measures that describe the interplay among different dimensions, for example whether a dimension "includes" another one, which and how many dimensions are "important" for the connectivity of a graph, what happens to local measures like the degree of a node when dealing with multiple dimensions, and so further.

In order to demonstrate the usefulness and wide applicability of the proposed framework, we implemented our proposed metrics, and performed a wide empirical analysis over many real and large multidimensional networks, that we acquired and prepared for this specific field experiment. We considered a large array of massive networks in diverse domains, ranging from query logs to social networks, customer networks, co-authorship bibliographic networks, the Web graph and a few more, and show how in each such case the introduced metrics - both the generalization of the known ones and the brand new multidimensional metrics - reveal a surprisingly high analytical power and suggest novel solutions to challenging real life problems.

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

The remainder of the paper is organized as follows: Section **??** presents the preliminary concepts needed for the comprehension of our work: the concept of multidimensional network together with its natural model, the multigraph, and the meaning of extracting analytical properties from them; Section **??** briefly overviews previous work related to our analysis; Section **??** gives a formal definition of the problem under investigation, and presents all the measures we define; Section **??** describes the datasets we use in the paper to evaluate our measures; in Section **??** we present our results obtained by applying the new measures on real data; finally in Section **??** we conclude our work giving also possible future research directions.

2 Background on Graphs and Network Analytics

In this section we recall some preliminary concepts which the reader may already be familiar with, but that constitute the basis of our work. We first provide a mathematical model of multidimensional networks, and then review the main properties and metrics for complex network analytics.

2.1 Multidimensional Networks as Multigraphs

The term *network* refers to the informal concept describing a structure composed of a set of elements and connections or interactions between them.

In a real network the entities may be connected by relations of different nature: for example, two persons may be linked because they are friends, colleagues, relatives or because they communicate to each other by phone, email, and so on. A network where a pair of entities may be linked by different kinds of links, having more than one connection between the two entites, is called a *multidimensional network*. We consider each possible type of relation between two entities a particular *dimension* of the network.

Often, the *graph* is used to model a network with its properties. In this graph, the entities are represented by nodes while a relation is modeled by an (directed or undirected) edge. In the case of a multidimensional setting a convenient way to model a network is hence a *labeled multigraph*. Intuitively, a labeled multigraph is a graph where both nodes and edges are labeled and where there can exist two or more edges between two nodes. Even though a labeled multigraph may be directed and undirected, for sake of simplicity, in our model we only consider undirected multigraphs, where edges have no direction. Moreover, in our context we do not consider node labels, thus we adopt a particular version of multigraph where only the edges are labeled. The model that we use in the remainder of the paper is hence the *edge-labeled undirected multigraph*. Formally, such a graph is 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., it is a set of triple of the form (u, v, l) where $u, v \in V$ are nodes and $l \in L$ is a label.

We assume that given a pair of nodes $u, v \in V$ and a label $l \in L$ it may exist only one edge (u, v, l). Thus, given $|L| = m$ each pair of nodes in G can be connected by at most m possible edges. In the following we denotes by $\mathcal{P}(L)$ the power set of the label collection L and by $\mathcal{G}$ the set of graphs of the form $G = (V, E, L)$.

When we use edge-labeled undirected multigraphs to model a multidimensional network, the set of nodes represents the set of entities or actors in the networks, the edges represent the interactions and relations between them and the edge labels describe the nature of the relations, i.e., the dimensions of the network. Given the strong correlation between labels and dimensions, in the following we use the term *dimension* in order to indicate *label*. Moreover, we denote by χ_E the characteristic function of E, which equals to 1 if a given edge (u, v, l) belonging to E, 0 otherwise. We also say that a node *belongs to* or *appears in* a given dimension l if it has at least one edge labeled with d. So, we define an operator $dim(v, l) : V \times L \rightarrow \{0, 1\}$ which equals to 1 if the node v appears in dimension d, 0 otherwise. Given a node $v \in V$, n_v is the number of dimensions in which v appears, i.e. $n_v = \mid \{l \in L \; s.t. \; dim(v, l) = 1\} \mid = \sum_{l \in L} dim(v, l)$. Similarly, given a pair of nodes $u, v \in V$, n_{uv} is the number of dimensions which label the edges between u and v, i.e., $n_{uv} = \mid \{l \in L \; s.t. \; \chi_E(u, v, l) = 1\} \mid = \sum_{l \in L} \chi_E(u, v, l)$.

2.2 Network Analytics

As said above, network data usually is modeled or represented by a graph. In literature, many analytical metrics have been defined in order to describe and analyze properties of a network. A metric is a function that can be defined on the whole network, or on the nodes, or on the edges. Defining meaningful metrics provides several advantages in the analysis of complex networks. From the simplest metric, the *degree* of a node, to more sophisticated ones, like the *betweenness centrality*, or the *eigenvector centrality* (a variant of which is behind the idea of Google's Pagerank [**?**]), several important results have been obtained in analyzing complex networks on real-world case studies. Section **??** briefly describes some of the most significant studies which involved the use of simple measures.

In graph theory, we have two categories of metrics: the global metrics describing characteristics of the whole network, and the local metrics describing local

properties either of a specific node or of a specific edge. The most important local metrics are:

Degree: The degree of a node v is the number of edges connected to v. It is well known that in real-world networks, the distribution of the degree is not random, but follows precise laws. The most common model of a network is the *scale-free* model, i.e. where the degree distribution follows a *power law* [**?**].

Neighborhood: The neighborhood of a node v in a graph is the set of all the nodes adjacent to v, that is not necessarily equal to the degree, since there may exist more than one edge between any two nodes in the case of a multigraph.

Shortest Path: The shortest path between two nodes in a network is the minimum of the lengths, in number of edges, of all the possible paths between those two nodes. Finding the shortest path is a well known problem in computer networks, and many routing algorithms rely on it [**?**].

Closeness Centrality: The closeness centrality is a measure of the centrality of a node within a network. The closeness centrality of a node v is defined as the mean length of the shortest paths between v and all other nodes reachable from v [**?**]. This measure of centrality can be easily used in information networks for applications such as Viral Marketing, where the goal is to find nodes that would maximize the spread of information, while minimizing the costs [**?**,**?**].

Betweenness Centrality: The betweenness centrality of a node v is defined as the number of shortest paths of the network that pass through v divided by the number of shortest paths of the network [**?**]. This is a good measures of the "resilience" of an information network to random or ad-hoc attacks: with a high variance of the betweenness centrality, in fact, we have that there are a few nodes very important for the flows of information, while there are many nodes with low importance. This means that the network is fairly resilient to random attacks, but very vulnerable to ad-hoc ones.

Edge Betweenness: The edge betweenness score of an edge measures the number of shortest paths that pass through it.

It is possible to define also some global metrics describing interesting aspects of a network, such as the mean degree, the average shortest path, etc. Moreover, some of them are calculated as "aggregates" of some local measure. The most popular global metrics defined on graphs are:

Connected Components: This function computes the number of connected components in a network. A connected component (or simply a component) is a maximal set of nodes such that, for any pair of the nodes in the set, there exists a connecting path. For example, if we consider the relationship among persons "working in the same company", this is likely to be modeled with a network consisting of many separated connected components (each being a *clique*).

Diameter: The diameter of a network is the length (in number of edges) of the longest shortest path between any two nodes. This measure gives an idea of how "wide" is a network, and it has been shown to be a very low value in many real networks [**?**].

Mean Shortest Path: The mean shortest path is the average length (in number of edges) of the shortest paths between any two nodes.

From the above list, it should be clear how important and powerful is to have a set of meaningful measures available for analyzing complex networks. When dealing with the multidimensional setting, the analysis scenario gets even richer, thanks to the availability of different dimensions to take into account. Therefore, many novel and meaningful metrics can be defined, which allow us to catch hidden dependencies or relationships among different dimensions. For instance, we may want to analyze the importance of a dimension with respect to another, the importance of a dimension for a specific node, and so on. In Section **??** we define several novel metrics that enable this kind of analyses. Moreover, we show how most of the metrics defined for classical monodimensional networks can be generalized in order to be applied to multidimensional networks.

3 Related work

In this section we give a summarization of the most relevant work related to our research, along two different perspectives. First, we provide a sketchy overview of the classical achievements of complex network analysis. Second, we assess the efforts towards multidimensional network analysis.

An exhaustive survey of network analytics concepts is provided by Newman [**?**], 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: the small world effect, the clustering coefficient, the degree distributions, the network resilience, together with various network generation models. The science of networks is today a highly visible fields, with brilliant books also tailored for broad dissemination [**?**,**?**,**?**].

More technically, 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 [**?**], Internet [**?**], the Web [**?**,**?**,**?**], citation networks [**?**], online social networks [**?**] and many others. Many large real-world networks exhibit particular laws of degree distribution, such as the power law (or

heavy-tailed) degree distribution. This property often comes together with the "small world" phenomenon, i.e., a relatively small average shortest path between any two nodes of the networks. In order to deal with outliers, the *effective* diameter [**?**] has been defined, found to be small for many real-world large networks [**?**,**?**,**?**,**?**]. Watts and Strogatz defined the clustering coefficient in [**?**], as a measure of the transitivity of the network. In some cases it is observed how the clustering coefficient decreases as the degree increases, a possible sign of a hierarchic network [**?**,**?**].

We mention another interesting survey paper by Chakrabarti and Faloutsos [**?**], where, besides the network properties, several properties of graph generators are analyzed. 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 the multidimensional networks perspective, there is little work so far on a general methodology for multidimensional network analysis, and relatively many works that address specific problems in a multidimensional setting.

The only paper in the first line that we aware of is [**?**], which introduces tha *OLAP graph*, a multidimensional view of graph data. The paper defines *informational* and *topological* dimensions over a graph. The first ones correspond simply to different observations of the same graph, while the second correspond to different hierarchical views of the graph. The paper presents a classification of measures on the OLAP graph, in terms of *distributive, algebraic* or *holistic*, depending on whether the measures of higher level cells can be easily computed from their lower level counterparts, without accessing base tuples. Finally, the paper presents possible optimization in the computation of some measure w.r.t. their distributive or holistic features. In summary, this work gives a multidimensional view of a graph to the purpose of defining the aggregation of different dimensions, but a systematic definition of analytical measures is missing. In particular, no new measures are defined, and the interplay among different dimensions is not investigated in any way. In other words, the OLAP graph is a method for supporting the navigation along the dimensions of a network, not a general framework for multidimensional network analysis.

On the other line, some recent works put emphasis on specific multidimensional social networks, such as, as an example, communication networks among people [**?**]. 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.

4 Multidimensional Analysis

The number of existent analytical properties of a graph can be extended to cope with multiple dimensions. Nevertheless, dealing with multiple dimensions raises questions on how such dimensions are correlated to each other, and if there are meaningful aggregate properties that require a specific definition, and that are meaningless in the monodimensional case.

In the following we denote by $D \subseteq L$ a set of dimensions of a network $G = (V, E, L)$ and by $D_v \subseteq L$ the set of dimensions where a specific node v appears.

4.1 Extending the Monodimensional Case

In this section we describe how the analytical measures defined on standard graphs (some of which were summarized in Section **??**) can be extended to deal with multiple dimensions. Moreover, we define new aggregate functions induced by some local or global measures.

In general, in order to adapt the classical metrics 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, we redefine some of the classical measures on graphs and networks, in order to follow the above approach. After this set of measures, we present the new measures we introduce on the multidimensional setting, that are meaningful only in this scenario.

Degree This function computes the degree of a node in a network, i.e. the number of edges connected to it. In order to cope with the multidimensional setting, we can define the degree of a node w.r.t a single dimension, w.r.t a set of dimensions and we can also analyze the average degree of a node within the network. To this end

Node	Degree(v,L)	Neighbors(v,L)	Degree(v,1)	Degree(v,2)	AVG_{Deg}(v,L)	Closeness(v,L)	TotSplit(v)	TotMix(v)
1	1	1	0	1	0.5	0.42	tt	ff
2	3	3	2	1	1.5	0.66	tt	ff
3	2	2	2	0	1	0.54	tt	ff
4	7	5	5	2	3.5	0.85	ff	ff
5	4	2	2	2	2	0.46	ff	tt
6	5	3	2	3	2.5	0.6	ff	ff
7	2	2	1	1	1	0.54	tt	ff
8	1	1	1	0	0.5	1	tt	ff
9	1	1	1	0	0.5	1	tt	ff
Node	$Neigh_{XOR1}$(v,L)	$Neigh_{XOR2}$(v,L)	DimRel(v,1)	$DimRel_W$(v,1)	$DimRel_{XOR}$(v,1)	DimRel(v,2)	$DimRel_W$(v,2)	$DimRel_{XOR}$(v,2)
1	0	1	0	0	0	1	1	1
2	2	1	0.66	0.66	0.66	0.33	0.33	0.33
3	2	0	1	1	1	0	0	0
4	3	0	1	0.8	0.6	0.4	0.2	0
5	0	0	1	0.5	0	1	0.5	0
6	0	1	0.66	0.33	0	1	0.66	0.33
7	1	1	0.5	0.5	0.5	0.5	0.5	0.5
8	1	0	1	1	1	0	0	0
9	1	0	1	1	1	0	0	0

Table 1 Summary of the values of the multidimensional metrics on the nodes of the toy example.

we have to redefine the domain of the classical degree function by including also the dimensions.

Definition 1 (Degree) Let $v \in V$ be a node of a network G. 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 between v and any other node labeled with one of the dimensions in D.

As it can be done for most of the measures that we present further, for this measure 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 all the measures below extending the multidimensional case, thus we avoid to repeat it for each of them.

Besides computing the average degree of the network, by summing all the degrees of the nodes and dividing by the number of nodes, we can also induce an aggregate function that computes the average of the degrees of a node v computed in different dimensions, by dividing for the number of dimensions considered.

Definition 2 (Average of the Degrees over dimensions) Let $v \in V$ be a node of a network G. The function $AvgDegree : V \times \mathcal{P}(L) \rightarrow \mathbb{R}$ defined as

$$AvgDegree(v, D) = \frac{Degree(v, D)}{|D|}$$

computes the average degree of a node v over the specific set of dimensions D of the network G. □

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

Example 1 Consider the multigraph in Figure **??** 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:

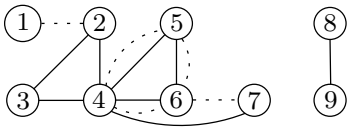


Fig. 2 A toy example. The solid line is dimension 1, the dashed line is dimension 2.

Measure	Global	Dim. 1	Dim. 2
ConnComponents	2(0)	3(1)	5(3)
Diameter	3	2	2
Net. Cluster Heterogeneity	0.33	/	/
NodeDimDegree	1	0.88	0.66
EdgeDimDegree	1	0.61	0.38
NodeDimDegreeUniq	1	0.37	0.16
EdgeDimDegreeUniq	1	0.62	0.4
NodeCorr	1	0.55	
EdgeCorr	1	0.3	
NodeParent	1	0.62	0.83
EdgeParent	1	0.37	0.6

Table 2 Values of the metrics on the dimensions of the toy example.

- $Degree(3, \{d_1\}) = 2$
- $Degree(3, \{d_2\}) = 0$
- $AvgDegree(3, \{d_1, d_2\}) = (2 + 0)/2 = 1$
- $AvgDegree(3, \{d_1\}) = 2/1 = 2$

Table **??** summarizes the values of all the measures computed on the nodes of the toy example.

In Section **??** we use the *Degree* function. We show that with this function it is possible to indentify dimensions that are good representative of the global degree distribution of the network, thus allowing tasks such as focused sampling or (lossy) compression of the graph. Also we show in Section **??** that even inside a multidimensional network that does not present a power law degree distribution there can be scale free dimensions.

Shortest Path As done for the degree, the classical shortest path definition has to be revisited in order to deal with the multidimensional setting, by extending the domain with a set of dimensions.

Definition 3 (Shortest Path) Let $u, v \in V$ be two nodes of a network G. The function $ShortestPath : V \times V \times \mathcal{P}(L) \rightarrow \mathbb{N}$ computes the length of the shortest

path (in terms of number of edges) between u and v, where the edges are labeled with dimensions in D. □

As in the classical definition, if are no paths between two nodes then the distance between them is ∞.

We also define the *Average Shortest Path* and the *Average of the Shortest Paths over dimensions*, which are two aggregate functions.

Definition 4 (Average Shortest Path) The function $ShortestPath_{AVG} : \mathcal{P}(L) \rightarrow \mathbb{R}$ is defined as

$$ShortestPath_{AVG}(D) = \frac{\sum_{\forall p \in SP_D} Length(p)}{|SP_D|}$$

where:

- SP_D denotes the set of shortest paths having only edges labeled with dimensions belonging to D, between node v and any node u reachable from it.
- $Length(p)$ denotes the length of the shortest path p in terms of number of edges.

It computes the average shortest path considering only the set of dimensions D. □

Definition 5 (Average of the Shortest Paths over dimensions) Let $u, v \in V$ be two nodes of a network G such that v is reachable from u. The function $AvgShortestPath : V \times V \times \mathcal{P}(L) \rightarrow \mathbb{R}$ defined as

$$AvgShortestPath(u, v, D) = \frac{\sum_{d \in D} ShortestPath(u, v, \{d\})}{|D|}$$

computes the average of the lengths of the shortest paths between two nodes u and v over the specific set of dimensions D of the network G. □

An interesting analysis that can be done when taking the dimensions into account in the definition of the shortest path is to verify the heterogeneity of the shortest path, i.e. verifying how many dimensions are traversed by a given shortest path. To this end, we define a function that computes the heterogeneity of a path.

Definition 6 (Path Heterogeneity) Let P be a path between two nodes of a multidimensional network, i.e. a sequence of labeled edges. The *Path Heterogeneity* function computes the ratio of dimensions in P with respect to the dimensions in the whole network. □

Given a set of paths it is possible to compute aggregate functions of the path heterogeneity measure, such as the average, the maximum and the minimum. In a multidimensional network, it is interesting to apply this measure on the set of the shortest paths: considering a transportation network, this would translate in knowing how many different trains, or tickets, a persons has to take to get to the destination. A possible variant of this is to count the number of "changes" of dimensions: even though the number of different crossed dimensions can be only two, it may happen that in order to go from one node to another one in the network, the shortest path is a sequence of $d_1 - d_2 - d_1 - ... - d_2$, which, in the transportation network, would mean to change train at every station. An interesting problem would be to modify Dijkstra's algorithm for computing the shortest path [**?**] to include also the change of edge label (i.e. dimension) as additional cost of the shortest path. This interesting analysis opens the way for many related problems and issues, and we plan in the future to investigate also in this direction.

Example 2 Continuing with the example of Figure **??** we have:

- the $ShortestPath(1, 7, \{d_1, d_2\}) = 3$ and its Heterogeneity is equal to 1, as this shortest path contains 2 edges of the dimension d_2 and 1 edge of the dimension d_1.
- $ShortestPath(1, 7, \{d_1\}) = \infty$
- $ShortestPath(6, 7, \{d_1, d_2\}) = 1$
- $ShortestPath(6, 7, \{d_1\}) = 2$
- $ShortestPath(6, 7, \{d_2\}) = 1$
- $AvgShortestPath(6, 7, \{d_1, d_2\}) = 1.5$
- $ShortestPath_{AVG}(\{d_1, d_2\}) = 1.6$

Closeness Centrality The closeness describes a particular kind of centrality of a node within a network, in terms of distance from all the other nodes. In the standard definition this measure is only defined on nodes. As done for the above measures, we modify the definition introducing the dimensions.

Definition 7 (Closeness Centrality) Let $v \in V$ be a node of a network G. The function $Closeness : V \times \mathcal{P}(L) \rightarrow [0, 1]$ is defined as

$$Closeness(v, D) = \frac{|\bar{V}|}{\sum_{u \in \bar{V}} ShortestPath(v, u, D)}$$

where $\bar{V}$ denotes the set of nodes reachable from v by a path, excluding v itself. □

Moreover, we define an aggregate function that computes the average of the closeness centrality computed over different dimensions.

Definition 8 (Average of the Closeness Centralities over Dimensions) Let $v \in V$ a node of a network G. The function $AvgCloseness : V \times \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$AvgCloseness(v, D_v) = \frac{\sum_{d \in D_v} Closeness(v, \{d\})}{|D_v|}$$

computes the average of the closeness centralities of a node v over the specific set of dimensions D_v of the network G. □

Please note that in this measure we explicitly indicate the set D_v, as it is not meaningful to consider the closeness in dimensions where the node does not appear.

Example 3 In the multidimensional network of the Figure **??**:

- if we consider all the dimensions and the node 7 we have six nodes reachable with a total number of 11 edges: $Closeness(7, \{d_1, d_2\}) = 6/11 = 0.54$
- if we consider only the dimension d_1 and the node 7 we have $Closeness(7, \{d_1\}) = 5/9 = 0.55$
- if we consider only the dimension d_2 and the node 7 we have $Closeness(7, \{d_2\}) = 3/5 = 0.6$
- the average of the closeness on the all the dimensions of the node 7 is $AvgCloseness(7, \{d_1, d_2\}) = (0.55 + 0.6)/2 = 0.57$

Betweenness Centrality While the closeness centrality takes into account the distance between a node and all the other nodes in a network, the betweenness centrality considers the number of shortest paths that passes through a node, thus emphasizing the analysis of the resilience of the network to the removal of important nodes. Also in this case, we would like to include a set of dimensions into account, hence we modify the standard definitions, introducing the followings.

Definition 9 (Betweenness Centrality) Let $v \in V$ be a node of a network G. The function $Betweenness : V \times \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$Betweenness(v, D) = \frac{\sum_{s,t \in V} SP_{svt}(D)}{SP_{st}(D)}$$

where:

- $SP_{svt}(D)$ denotes the number of shortest path between the nodes s and t passing through v, only considering edges belonging to the set of dimensions D
- $SP_{st}(D)$ denotes the number of shortest path between the nodes s and t, considering only considering edges belonging to the set of dimensions D. □

We can also define an aggregate function that computes the average of the betweenness centralities computed on different dimensions.

Definition 10 (Average of the Betweenness Centralities over Dimensions) Let $v \in V$ be a node of a network G. The function $AvgBetweenness : V \times \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$AvgBetweenness(v, D) = \frac{\sum_{d \in D} Betweenness(v, \{d\})}{|D|}$$

computes the average of the betweenness centralities of a node v computed over the specific set of dimensions D of the network G. □

Connected Components Detecting the connected components in a network, specially in social networks, is a powerful as easily computable way to better understand the topology of the network. In the following we compute the number of connected components of a multidimensional network, including also a set of dimensions into account.

Definition 11 (Connected Components) The function $CC : \mathcal{G} \times \mathcal{P}(L) \rightarrow \mathbb{N}$, called *Connected Components*, computes the number of connected component of a graph considering only the edges labeled with dimensions included in a given set D. □

Definition 12 (Average of the Connected Components over Dimensions) The function $AvgCC : \mathcal{G} \times \mathcal{P}(L) \rightarrow \mathbb{R}$, called *Average of the Connected Components over Dimensions*, is defined as

$$AvgCC(G, D) = \frac{\sum_{d \in D} CC(v, \{d\})}{|D|}$$

and computes the average of the number of connected components over the specific set of dimensions D of a given network G. □

Example 4 Considering the multidimensional network of the Figure **??**:

- if we consider only the dimension d_1 we have 3 components, as we can consider node 1 as a component composed by only one node
- if we consider only the dimension d_2 we have 5 components, as we can consider nodes 3, 8 and 9 as above
- if we consider all the dimensions of the network we have 2 components
- the connected component average of the network are $(3 + 5)/2 = 4$

In Section **??** we perform an analysis based on the statistics on the connected components by looking at different sets of dimensions of two networks. We show how different is the change of the structure of the connected components and the giant component by adding dimensions in networks that indeed have a similar structure in the final aggregate.

Diameter The classical definition of the diameter is the length of the longest shortest path between any pair of nodes in the network. Having re-defined the notion of "shortest path" we can define the concept of diameter in terms of it.

Definition 13 (Diameter) The function $Diameter : \mathcal{G} \times \mathcal{P}(L) \rightarrow \mathbb{N}$ computes the length of the longest shortest path of a network G considering only edges labeled with dimension belonging to a specific set D. □

Clearly, on the diameter it is possible to define aggregate functions as for the shortest path. It is also interesting to measure the difference between the diameter computed considering a given set of dimension and the diameter of the whole network.

Example 5 On Figure **??**:

- $Diameter(G, L) = 3$
- $Diameter(G, \{d_1\}) = 2$

4.2 Leveraging the Multidimensional Case

In the following, we build the most important part of our theory: we define new metrics that are meaningful only in the multidimensional setting, as they try to analyze the true interplay that occurs among them, rather than simply calculating a measure on an aggregate of a set of dimensions. As we have done above, we comment our definitions with their values on the toy example, to better clarify their meaning.

Neighbors When considering the multidimensional setting, it is straightforward to see that the degree notion becomes not complete, as the number of edges adjacent to a node and the number of neighbors of the node itself are not related anymore. In order to overcome to this problem, we define a new measure computed on a node. First, we present the formal definition of neighbors of a node given a set of dimensions. In the paper we use the expression *directly reachable* to indicate a node for which the length of the shortest path from a given node is 1.

Definition 14 (Neighbor Set) Let $v \in V$ be a node and a set of dimensions of a network G. The function $NeighborSet : V \times \mathcal{P}(L) \rightarrow \mathcal{P}(V)$ defined as

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

computes the collection of all the nodes directly reachable from a node v following only edges labeled with dimensions in D. □

At this point, as the degree counts the number of adjacent edges, we would like to count the number of adjacent nodes.

Definition 15 (Neighbors) Let $v \in V$ be a node of a network G. The function $Neighbors : V \times \mathcal{P}(L) \rightarrow \mathbb{N}$ is defined as

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

□

Note that in the monodimensional case, this measure corresponds to the degree.

In the following, we give also a variant of the above definition, that takes into account only neighbors reachable with edges belonging only to a specific set of dimensions, excluding the ones that can be reached by edges belonging to other dimensions.

Definition 16 (Neighbors$_{XOR}$) Let $v \in V$ be a node of a network G. The function $Neighbors_{XOR} : V \times \mathcal{P}(L) \rightarrow \mathbb{N}$, defined as

$$Neighbors_{XOR}(v, D) = \sum_{u \in V} k_{uv}(D)$$

where:

$$k_{uv}(D) = \begin{cases} 1 & if \; \forall (u, v, d) \in E : d \in D \\ 0 & otherwise \end{cases}$$

computes the number of neighboring nodes reachable following only edges labeled with dimensions in D, and not reachable following edges labeled with other dimensions. □

As usual, we define the aggregate function that computes the average on the numbers of neighbors.

Definition 17 (Average of Neighbors over Dimensions) Let $v \in V$ be a node of a network G. The average of the neighbors of v with respect to a set of dimensions is the function $AvgNeighbors : V \times \mathcal{P}(L) \rightarrow \mathbb{R}$ defined as

$$AvgNeighbors(v, D) = \frac{\sum_{d \in D} Neighbors(v, \{d\})}{|D|}.$$

□

Example 6 In Figure **??**, the followings hold:

- $Neighbors(4, \{d_1, d_2\}) = 5$
- $Neighbors(4, \{d_1\}) = 5$
- $Neighbors(4, \{d_2\}) = 2$
- $Neighbors_{XOR}(4, \{d_1\}) = 3$
- $Neighbors_{XOR}(4, \{d_2\}) = 0$
- $AvgNeighbors(4, \{d_1, d_2\}) = (5 + 2)/2 = 3.5$
- $AvgNeighbors(3, \{d_1, d_2\}) = (2 + 0)/2 = 1$
- $AvgNeighbors(3, \{d_1\}) = 2/1 = 2$

In Section **??** we investigate some aspects of the neighbor distribution. We show that this distribution generally follows the same behavior of the degree distribution, thus presenting a power law in many cases.

The following two new concepts, meaningful only in the multidimensional case, help in understanding the interplay among dimensions from two different points of view: the node and the dimension.

Node Relevance In multidimensional networks it is interesting to measure how much a particular node is important, w.r.t the connectivity of the network, to a set of dimension. To this end we define the novel notion of *node relevance*.

Definition 18 (Node Relevance) Let $v \in V$ be a node of a network G. The function $NodeRelevance : V \times \mathcal{P}(L) \rightarrow \mathbb{N}$ is defined as the number of dimensions of D where the average shortest path $ShortestPath_{AVG}(D)$ increases after v is removed from the network. □

Example 7 In the Figure **??** we are in the particular situation in which the Node Relevance is equal to zero for all the nodes, i.e. if we remove a node the average shortest path does not increase in any dimension.

Dimension Relevance From the opposite view, another interesting question is how much is important a particular dimension over the others for the connectivity of a node. In order to answer this question, we define the concept of *dimension relevance*. Intuitively, it analyzes how that node gets disconnected if we remove that dimension from the network.

Definition 19 (Dimension Relevance) Let $v \in V$ be a node of a network G. The function $DimRelevance : V \times \mathcal{P}(L) \rightarrow [0,1]$ defined as

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

computes the ratio of neighbors directly reachable from node v following edges belonging to dimensions in D.□

In the following, we also define two variants of the Dimension Relevance: the first one allows the neighbors to be reached only by edges belonging to a specific set of dimensions, while the second is a weighted version of it, that takes into account also the number of alternatives (i.e. the number of edges belonging to dimensions not included in the given set) for reaching a node. Further, we state a theorem that relates the values of the three definitions.

Definition 20 (Dimension Relevance *XOR*) Let $v \in V$ be a node of a network G. The *Dimension Relevance XOR* is the function $DimRelevance_{XOR} : V \times \mathcal{P}(L) \rightarrow [0,1]$ defined as

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

computing the fraction of neighbors directly reachable from node v following edges belonging only to dimensions included in D. □

Definition 21 (Weighted Dimension Relevance) Let $v \in V$ be a node of a network $G = (V, E, L)$. The function $DimRelevance_W : V \times L \rightarrow [0,1]$, called *Weighted Dimension Relevance*, is defined as

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

where:

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

Theorem 1 *Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a multidimensional network $G = (V, E, L)$, respectively. Then we have that*

$$DimRelevance_{XOR}(v, D) \leq DimRelevance_W(v, D) \leq DimRelevance(v, D).$$

Proof In order to prove this theorem it is sufficient to show that

$$Neighbors_{XOR}(v, D) \leq \sum_{u \in NeighborSet(v,D)} \frac{n_{uvd}}{n_{uv}} \quad (1)$$

and

$$\sum_{u \in NeighborSet(v,D)} \frac{n_{uvd}}{n_{uv}} \leq Neighbor(v, D) \quad (2)$$

as $DimRelevance_{XOR}(v, D)$, $DimRelevance_W(v, D)$ and $DimRelevance(v, D)$ have the same denominator. Suppose

$A = Neighbors_{XOR}(v, D)$
$B = \sum_{u \in NeighborSet(v,D)} \frac{n_{uvd}}{n_{uv}}$
$C = Neighbors(v, D)$.

First of all, we prove the inequality (1). If the node v is connected to a neighbor u only by edges labeled with dimensions in D then in both the formulas, A and B, u contributes with 1; if they are connected only by edges labeled with dimensions that do not belong to D then in both the formulas, A and B, u contributes with 0; finally, if they are connected by some edges labeled with dimensions in D and some edges labeled with dimensions that do not belong to D then in A the node u contributes with a value equal to 0 while in B it contributes with a value greater than 0. So, we have that $A \leq B$.

Now, we prove the inequality (2). If the node v is connected to a neighbor u only labeled with dimensions in D then in both the formula B and C it contributes with 1; if they are connected only by edges labeled with dimensions that do not belong to D then in A and B u contributes with 0; finally, if they are connected by some edges labeled with dimensions that do not belong to D

and some edges labeled with dimensions in D then in B the node u contributes with a value equal to $\frac{n_{uvd}}{n_{uv}} < 1$ ($d \in D$) while in C it contributes with 1. So, we have that $B \leq C$. □

Example 8 Considering Figure **??**, the followings hold:

- $DimRelevance(4, \{d_1, d_2\}) = 1$
- $DimRelevance(4, \{d_1\}) = 1$
- $DimRelevance(4, \{d_2\}) = 2/5 = 0.4$
- $DimRelevance_W(4, \{d_1\}) = (1 + 1 + 0.5 + 0.5 + 1)/5 = 0.8$
- $DimRelevance_W(4, \{d_2\}) = (0 + 0 + 0.5 + 0.5 + 0)/5 = 0.2$
- $DimRelevance_{XOR}(4, \{d_1\}) = 3/5 = 0.6$
- $DimRelevance_{XOR}(4, \{d_2\}) = 0/5 = 0$

In Section **??** we investigate the distributions of the three variants of dimension relevance. We show how it is possible to infer some pieces of knowledge about the structure of the network and the interplay among dimensions by only looking at the global dimension relevance distribution.

Totally Split and Totally Mixed We introduce two novel notions on the nodes of a multidimensional network: *Totally Split* and *Totally Mixed*. They are derived from the combination of the functions Degree and Neighbors. Intuitively, these measures in some way describe the structure around a given node in terms of edge density: if the node is totally split this structure is sparse, while if the node is totally mixed it is dense and redundant.

Definition 22 (Totally Split) A node $v \in V$ is called *Totally Split* 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 totally split we have

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

Definition 23 (Totally Mixed) A node $v \in V$ is called *Totally Mixed* 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 totally mixed we have

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

Example 9 In Figure **??** we have several Totally Split 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 Totally Mixed node: node number 5 is connected via both the dimensions with each of its neighbors.

Lemma 1 *Let $v \in V$ be a node of a multidimensional network G. Let D_v the set of dimensions where v appears. Then we have*

$$Neighbors(v, L) = Neighbors(v, D_v) = Neighbors_{XOR}(v, D_v).$$

Proof First of all, we show that $Neighbors(v, L) = Neighbors(v, D_v)$. By definition, $Neighbors(v, L) = |\{u \in V | (u, v, d) \in E \wedge \; d \in L\}|$ and $Neighbors(v, D_v) = |\{u \in V | (u, v, d) \in E \wedge \; d \in D_v\}|$. We can say that $Neighbors(v, L) = |\{u \in V | (u, v, d) \in E \wedge \; d \in L\}| = |\{u \in V | (u, v, d) \in E \wedge \; d \in D_v\} \cup \{u \in V | (u, v, d) \in E \wedge \; d \in (L \setminus D_v)\}|$. In this last union, the second set is empty since it contains all the nodes not connected with v, thus we can conclude that $Neighbors(v, L) = Neighbors(v, D_v)$. Now, we prove that $Neighbors(v, D_v) = Neighbors_{XOR}(v, D_v)$. First, it is easy to notice that $Neighbors_{XOR}(v, D_v) = |\{u \in V | (u, v, d) \in E \wedge \; d \in D_v\} \setminus \{u \in V | (u, v, d) \in E \wedge d \in (L \setminus D_v)\}|$. The set $\{u \in V | (u, v, d) \in E \wedge d \in (L \setminus D_v)\}$ is empty since it contains all the nodes not connected with v, so we can conclude that $Neighbors(v, D_v) = Neighbors_{XOR}(v, D_v)$. □

We can state a few interesting properties of the totally split and totally mixed nodes.

Theorem 2 *Let $v \in V$ be a node of a multidimensional network G. Let D_v the set of dimensions where v appears.*

(a) The node v is Totally Mixed if and only if

$$\forall d \in L : \; DimRelevance_W(v, \{d\}) = \tfrac{1}{|L|}.$$

(b) The node v is Totally Mixed if and only if

$$\forall d \in L : \; DimRelevance(v, \{d\}) = 1$$

(c) The node v is Totally Split if and only if

$$\textstyle\sum_{d \in D_v} DimRelevance_{XOR}(v, \{d\}) = 1$$

(d) If the node v is Totally Mixed then

$$\textstyle\sum_{d \in D_v} DimRelevance_{XOR}(v, \{d\}) = 0$$

Proof In the following we prove the four points of the theorem.

part (a) ($\Rightarrow$) Assume that the node v is totally mixed, thus for each node $u \in NeighborsSet(v, \{d\})$ there are $|L|$ edges between v and u. This means that in $DimRelevance_W(v, \{d\})$ each term $n_{uv} = |L|$, therefore it is immediate to conclude that

$DimRelevance_W(v, \{d\}) = \frac{1}{|L|}$.

($\Leftarrow$) Assume that $DimRelevance_W(v, \{d\}) = \frac{1}{|L|}$. In order to show that v is totally mixed it is sufficient to show that in $DimRelevance_W(v, \{d\})$ for each term we have $\frac{1}{n_{uv}} = \frac{1}{|L|}$. We know that:

(1) for each $u \in NeighborsSet(v, L)$ it exists at least a labeled edge;
(2) in $DimRelevance_W(v, \{d\})$ we have $Neighbors(v, L)$ terms;
(3) $|L| \geq n_{uv}$.

Since, by hypothesis, $DimRelevance_W(v, \{d\}) = \frac{1}{|L|}$, we have $|L| = n_{uv}$, and this means that v is totally mixed.

part (b) By definition of totally mixed node and *DimRelevance*.

part (c) ($\Rightarrow$) Assume that the node v is totally split. By definition, we have

$\sum_{d \in D_v} DimRelevance_{XOR}(v, \{d\}) = \frac{\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\})}{Neighbors(v, L)}$

Therefore, we have to show that

$\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\}) = Neighbors(v, L)$.

By the Lemma **??**, $Neighbors(v, L) = Neighbors(v, D_v)$. By definition, $Neighbors(v, D_v) = |\{u \in V \ s.t. \ (u, v, d) \in E \ \wedge \ d \in D_v\}| = |\bigcup_{d \in D_v} \{u \in V \ s.t. \ (u, v, d) \in E\}|$. By hypothesis, v is totally split, hence v is connected to each neighbor u by only one edges, and this means that $\bigcup_{d \in D_v} \{u \in V \ s.t. \ (u, v, d) \in E\}$ is a disjoint union and so we conclude that $\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\}) = |\bigcup_{d \in D_v} \{u \in V \ s.t.$
$(u, v, d) \in E\}| = Neighbors(v, L)$.
($\Leftarrow$) Suppose that $\sum_{d \in D_v} DimRelevance_{XOR}(v, \{d\}) = 1$, i.e., $\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\}) = Neighbors(v, L)$. By the Lemma **??**, $Neighbors(v, L) = Neighbors(v, D_v)$ $= |\{u \in V \ s.t. \ (u, v, d) \in E \ \wedge \ d \in D_v\}| = |\bigcup_{d \in D_v} \{u \in V \ s.t. \ (u, v, d) \in E\}|$. So, $\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\})$ $= |\bigcup_{d \in D_v} \{u \in V \ s.t. \ (u, v, d) \in E\}|$. This is possible only if $\bigcup_{d \in D_v} \{u \in V \ s.t. \ (u, v, d) \in E\}$ is a disjoint union and so we conclude that v is a totally split node.

part (d) Assume that the node v is totally mixed, thus it is connected to each neighbor by $|L|$ edges. By definition, we have

$\sum_{d \in D_v} DimRelevance_{XOR}(v, \{d\}) = \frac{\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\})}{Neighbors(v, L)}$

Therefore, it is sufficient to show that

$\sum_{d \in D_v} Neighbors_{XOR}(v, \{d\}) = 0$.

But this is straightforward since, by hypothesis, v is totally mixed. □

Now, having introduced the definitions of totally split and totally mixed node, we can state a theorem on the relation between the degree and the number of neighbors.

Theorem 3 *Let $v \in V$ be a node of a multidimensional network $G = (V, E, L)$ and D_v the set of dimension where v appears. Then we have that*

$$AvgDegree(v, D_v) \leq Neighbors(v, L) \leq Degree(v, L).$$

Proof In order to prove the two inequalities we consider the extreme cases, i.e., when v is totally mixed and when v is totally split. First of all, we show that

$$AvgDegree(v, D_v) \leq Neighbors(v, L).$$

By definition we have $AvgDegree(v, D_v) = \frac{Degree(v, D_v)}{D_v}$. When v is totally mixed we have that $Degree(v, D_v) = |L| \times Neighbors(v, L)$, thus

$$AvgDegree(v, D_v) = Neighbors(v, L) \quad (1)$$
$$Neighbors(v, L) \leq Degree(v, D_v) \quad (2)$$

When v is totally split we have

$$Degree(v, L) = Neighbors(v, L). \quad (3)$$

Moreover, since $Degree(v, L) = Degree(v, D_v)$ we have

$$AvgDegree(v, D_v) \leq Neighbors(v, L). \quad (4)$$

Considering (1) and (4) we can conclude that the first inequality is true; by (2) and (3) we also conclude that the second inequality is true. □

In Section **??** we present some examples of actual Totally Split and Totally Mixed nodes. We show how this property can come in help in real problems like improving the quality of a search engine.

Cluster Heterogeneity Like other metrics, also the clustering coefficient can be straightforwardly computed following its classical definition on a single dimension or on a set of dimensions, and we omit these definitions. But there is an interesting generalization of this coefficient when we are dealing with multidimensional networks: the *Cluster Heterogeneity*. Given two or more dimensions, this measure computes how many edges are the only connection between any two nodes in a *multidimensional triangle*. In the following, first of all we define what is a multidimensional triangle, then we introduce the notion of cluster heterogeneity for a single multidimensional triangle and finally we define the notion of *network clustering heterogeneity*.

Definition 24 (Multidimensional Triangle) A *multidimensional triangle* is a clique of three nodes in which at least one edge belongs to a different dimension than another edge in the clique.

Definition 25 (Clustering Heterogeneity) The *Clustering Heterogeneity* (denoted as ch) of a multidimensional triangle t is defined as

$$ch(t) = \frac{sc}{3}$$

where sc denotes the number of edges that are the sole connection between two nodes in the multidimensional triangle t. □

While the above is a local measure computed on a triangle, it is possible to define a global measure on the network.

Definition 26 (Network Clustering Heterogeneity) The *Network Clustering Heterogeneity* of a network G is defined as:

$$NetCh(D) = \frac{\sum_{t \in G'} ch(t)}{\mid \{t \mid t \in G'\} \mid}$$

where:

- G' denotes the triple (V', E', D) in which $V' \subseteq V$ contains only the nodes belonging to the dimensions in D and $E' \subseteq E$ only contains the edges belonging to dimension in D.
- t denotes a multidimensional triangle.

□

Example 10 In Figure **??** the whole network has three triangles. Only two of them are multidimensional: 4, 5, 6 and 4, 6, 7. These two triangles contain only two pairs of nodes that are connected by only one dimension. Thus the Network Clustering Heterogeneity is equal to $2/(3*2) = 0.33$. The Clustering Heterogeneity for triangle 4, 5, 6 is zero and for triangle 4, 6, 7 is 0.66.

In Section **??** we perform a clustering heterogeneity analysis. We show how this metric can help to infer some considerations in the problem of investigating the information propagation.

Dimension Degree An interesting analysis on multidimensional networks is to understand which percentage of nodes or edges are contained in a specific dimension. To this aim we define a novel measure called *Dimension Degree*.

Definition 27 (Node Dimension Degree) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $DimDegree_{Node} : L \rightarrow [0, 1]$ defined as

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

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

Definition 28 (Edge Dimension Degree) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $DimDegree_{Edge} : L \rightarrow [0, 1]$ defined as

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

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

We also introduce another measure called *Dimension Degree Uniqueness*, which takes into account the number of nodes or edges that belongs *only* to one dimension.

Definition 29 (Node Dimension Degree Uniqueness) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $ddu_{node} : L \rightarrow [0, 1]$ defined as

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

computes the ratio of nodes that belong only to the dimension d. □

Definition 30 (Edge Dimension Degree Uniqueness) Let $d \in L$ be a dimension of a network $G = (V, E, L)$. The function $ddu_{edge} : L \rightarrow [0, 1]$ defined as

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

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 11 In Figure **??** the Edge Dimension Degree of dimension d_1 is 0.61 since it has 8 edges out of the 13 total edges of the network. Its Edge Dimension Degree Uniqueness is equal to $5/8 = 0.625$. The Node Dimension Degree for the same dimension d_1 is 0.88 (8 nodes out of 9) and its Node Dimension Degree Uniqueness is 0.375 (3 unique nodes out of 8).

In Section **??** we show how these metrics can be used in order to perform a temporal analysis of a social network in which the dimensions are defined as temporal snapshots.

Dimension Correlation The following two measures are among the most important when detecting the interplay among dimensions. Intuitively, they give an idea of how redundant are two dimensions, if we can expect two nodes to be connected by a given dimension when they are found to be connected by a specific one, and so on.

Definition 31 (Node Correlation) Let $d_1, d_2 \in L$ be two dimensions of a network $G = (V, E, L)$. The *Node Correlation* is the function $\rho_{node} : L \times L \rightarrow [0, 1]$ defined as

$$\rho_{node}(d_1, d_2) = \frac{|V_{d_1} \cap V_{d_2}|}{|V_{d_1} \cup V_{d_2}|}$$

where V_{d_1} and V_{d_2} denote the nodes belonging to dimensions d_1 and d_2, respectively. It computes the ratio of nodes belonging to both the dimensions over the total number of nodes belonging to at least one of them. □

Definition 32 (Edge Correlation) Let $d_1, d_2 \in L$ be two dimensions of a network $G = (V, E, L)$. The *Edge Correlation* is the function $\rho_{edge} : L \times L \rightarrow [0, 1]$ defined as

$$\rho_{edge}(d_1, d_2) = \frac{|E_{d_1} \cap E_{d_2}|}{|E_{d_1} \cup E_{d_2}|}$$

where E_{d_1} and E_{d_2} denote the edges belonging to dimensions d_1 and d_2, respectively. It computes the ratio of edges belonging to both the dimensions over the total number of edges belonging to at least one of them. □

Definition 33 (Dimension Correlation Average) The *Dimension Correlation Average* is the function $Avg_{\bar{\rho}} : \mathcal{P}(L) \rightarrow [0, 1]$ defined as

$$Avg_{\bar{\rho}}(D) = \frac{\sum_{d_1, d_2 \in\ L} \bar{\rho}(d_1, d_2)}{|D|}, \qquad d_1 \neq d_2$$

where $\bar{\rho}$ can be either the node correlation function ρ_{node} or the edge correlation function ρ_{edge}. □

Intuitively, it computes the overall correlation in the whole network, giving an idea of the total redundancy residing in the different dimensions of the network, either in terms of nodes (entities), or in terms of edges (relations).

Example 12 In Figure **??**, the Node Correlation finds as common nodes of the two dimensions: 2, 4, 5, 6 and 7. Thus its value is equal to $5/9 = 0.55$. The Edge Correlation, on the other side, is equal to $3/10 = 0.3$.

In Section **??** we show how the dimension correlation can be used for the detection of *hierarchies* in networks, sometimes called *multilevel* analysis [**?**].

Parent Besides the correlation, another important kind of relationship between two dimensions is whether one "includes" another one, or, as we say, it is its *parent*. This measure finds the implicit hierarchy among different dimensions.

Definition 34 (Node Parent) Let $d_1, d_2 \in L$ be two dimensions of a network $G = (V, E, L)$. The *Node Parent* is the function $NodeParent : L \times L \rightarrow [0, 1]$ defined as

$$NodeParent(d_1, d_2) = \frac{|V_{d_1} \cap V_{d_2}|}{|V_{d_1}|}$$

where V_{d_1} and V_{d_2} denote the nodes belonging to dimensions d_1 and d_2, respectively. It computes the ratio of nodes belonging to d_1 that belong also to d_2. □

Definition 35 (Edge Parent) Let $d_1, d_2 \in L$ be two dimensions of a network $G = (V, E, L)$. The *Edge Parent* is the function $EdgeParent : L \times L \rightarrow [0, 1]$ defined as

$$EdgeParent(d_1, d_2) = \frac{|E_{d_1} \cap E_{d_2}|}{|E_{d_1}|}$$

where E_{d_1} and E_{d_2} denote the edges belonging to dimensions d_1 and d_2, respectively. It computes the ratio of edges belonging to d_1 that belong also to d_2. □

Example 13 In Figure **??** we have:

- $NodeParent(d_1, d_2) = 5/8 = 0.625$
- $NodeParent(d_2, d_1) = 5/6 = 0.83$
- $EdgeParent(d_1, d_2) = 3/8 = 0.375$
- $EdgeParent(d_2, d_1) = 3/5 = 0.6$

In Section **??** we show that the parent metrics can have useful applications in marketing, by highlighting some possible unknown relationships between products and costumers.

Dimension Closeness Similarly to the node case, we introduce a notion of distance between a dimension and all the others, called *Dimension Closeness*. In order to define it, first we introduce the definition of distance between two dimensions.

Definition 36 (Dimension Distance) Let $d_1, d_2 \in L$ be two dimensions of a network $G = (V, E, L)$. The function $DimDistance : L \times L \rightarrow \mathbb{N}$, called *Dimension Distance*, is defined as

$$DimDistance(d_1, d_2) = \frac{\sum_{p \in P} Length(p)}{|P|}$$

where:

- P denotes the set of paths from any node v belonging to the dimension d_1 and any node belonging to d_2 and reachable from v
- $Length(p)$ denotes the length of the path p in terms of number of edges. □

Now, using the Definition **??** we can define the *Dimension Closeness*.

Definition 37 (Dimension Closeness) Let $d_1 \in L$ be a dimension of a network $G = (V, E, L)$. The function $DimCloseness : L \times L \rightarrow [0, 1]$, called *Dimension Closeness*, is defined as

$$DimCloseness(d_1) = \frac{1}{\sum_{d \in L, d \neq d_1} DimDistance(d, d_1)}.$$

□

In the remainder of the paper, we present our experimental analysis, where we applied most of the above definitions to several large real-world networks. For some of the analysis we give suggestions for their practical usage in real life problems, such as compression, temporal analysis, computational advertisement, and a few others.

5 Datasets and Tools

In this subsection we present the data used in this paper, and the tools used to compute the statistics.

5.1 Data

For our analysis we used different kinds of real-world networks: a social network, two email datasets, the data coming from a large Italian chain of retail distribution, a query log dataset, a crawl of the Uk Web graph and the well known bibliographic dataset DBLP. For each of them, after a brief description of the collection, we present the pre-processing stages needed to extract the networks, which dimensions we modeled and a few rough statistics. Table **??** summarizes the following list, together with more complete statistics.

Query logs[1]**.** This data consists of approximately 20 millions of queries submitted by 650.000 users from March to May in 2006 to the America On Line search engine and was well described in [**?**]. A record on this query log represents the visit to a result for a query or the submission of a query (if no result is visited). Each record stores an anonymous ID that allows to group queries from the same user without revealing the AOL users nickname, the query submitted by the user, the date and hour of the submission of the query, the rank position of the result visited by the user on each record and the domain portion of the URL of the result visited.
From this dataset, we extracted a word-word network of query terms, consisting in roughly 200k words (nodes), after removing a list of stop-words (words too generic and frequent like articles, prepositions, punctuation and so on). We connected two words if they appeared together in a query, ending up with roughly 3M of edges. As dimensions we used the rank positions of the visited results, grouped in 6 equi-populated bins: 1 for rank 1, 2 for ranks 2-3, 3 for ranks 4-6, 4 for ranks 7-10, 5 for ranks 11-58, 6 for ranks 59-500.

Flickr[2]**.** This dataset comes from the well known photo sharing service, by crawling the data via the available APIs[3]. Part of the data was obtained from the HPC-Lab[4] group of the ISTI-CNR in Pisa, which already crawled the information about 106M of pictures and was described and used in [**?**].
As the service is really powerful, we were able to extract implicit and explicit dimensions of the social network residing in this data. From those pictures, in fact, we extracted the list of all the users related to them and among 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 contact. For roughly 1.3M users we obtained about 1G edges, spread on the mentioned four dimensions.

Enron[5]**.** This archive contains 619,446 email messages complete with senders, recipients, cc, bcc, and text sent and received from 158 Enron's employees [**?**]. We took from the entire dataset the "from", "to", "cc", "bcc", "subject" and "date" fields in each email in the "sent" folder of every employee. We took only the emails that were sent to other Enron employees, removing the outgoing emails. We also performed basic cleaning by removing emails with empty subjects, noise, and so on. After the cleaning stage, the number of remaining emails was about 12k. The method for creating the network is trivial: two users are linked if they have exchanged at least one email. We identified 7 possible dimensions, corresponding to the days of the week: an edge belongs to the "Monday" dimension if at least one email was exchanged on Monday between the two connected users. We ended up with roughly 6k nodes (employees) and about 30k edges.

Newsgroups[6]**.** This dataset consists of Usenet articles collected from 20 different newsgroups about general discussions on politics and religion, technical discussions on computers and hardware, general discussions on hobbies and so on. It was first described and used in [**?**,**?**]. Over a period of time, 1000 articles were taken from each of the newsgroups, which

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

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

[3] http://www.flickr.com/services/api

[4] http://hpc.isti.cnr.it

[5] http://www.cs.cmu.edu/~enron

[6] http://people.csail.mit.edu/jrennie/20Newsgroups

Dataset	Nodes	Dimensions	#Dim	#Nodes	#Edges	Avg Deg	Density
Query logs	words	rank bins	6	184,760	3,565,820	38.58	$3.48e^{-5}$
Flickr	users	friend, favorite, tag, comment	4	1,186,895	922,237,122	1554.03	$3.27e^{-4}$
Enron	users	email exchange same day	7	5,912	30,893	10.46	$2.52e^{-4}$
Newsgroups	users	email reply same day/same newsgroup	27	5,897	201,998	68.5	$4.3e^{-4}$
Supermarket	customers	same dept./class of item $\geq$ 10 times	27	5,137	3,834,497	1492.9	$1.08e^{-2}$
WebUk	web pages	link found in a monthly crawl	12	133,633,040	42,059,385,177	629.48	$3.93e^{-7}$
Dblp	authors	paper together same year	29	582,201	2,633,249	9.04	$5.36e^{-7}$

Table 3 Summary of the datasets used, with some statistics. Please refer to the text for a complete description of this Table.

makes an overall number of 20,000 documents in this collection. We took from each sent email the "from", "to" and "date" field. After a cleaning stage, the number of remaining emails was about 18,000. In this case we created an edge between two users if both have sent to a newsgroup a message with the same subject. As in the Enron dataset, even in this case we identified seven dimensions, following the choice of the day of the week. We also defined a separate dimension for each Newsgroup (an edge belongs to the dimension of the newsgroups to which the user sent the message), thus defining a total number of 27 dimensions.

The basic statistics about this data are a total of roughly 6k nodes (users), connected by roughly 200k edges spread on 27 dimensions (days of the week, or name of the newsgroup).

Supermarket. This data comes from one of the largest Italian chain of supermarkets. The data was collected in two years of purchases in shops located in the western Italian coast. For every store, we had the complete data in each receipt issued for every purchase by the cashiers. The incredible amount and richness of available data was subjected to a strong stage of preprocessing and selection, in which we decided to take only one month of purchases (December 2008), from only six cities (Avellino, Naples, Leghorn, Grosseto, Rome). We ended up with a total amount of 7.5M of transactions.

At the end of our preprocessing, we came to a dataset containing about 5k nodes (customers), connected by roughly 4M edges, where there is an edge if two users bought the same product at least 10 times. Each edge is labeled with the marketing department (such as "container", "self service", "bread") or class (such as "packaged for sale", "fresh", "very fresh") associated to the item, for a total of 27 dimensions.

WebUk[7]**.** This network consists of 12 monthly snapshots of a crawl of the Uk Web graph [**?**] available via the WebGraph framework [**?**], taken from June 2006 to May 2007, and consisting of about 133k nodes and 5G multi-labeled edges, where for each edge there are 12 bits representing the presence of that edge in one of the 12 snapshot.

We naturally considered each snapshot as a dimension of the graph, and we duplicated the edges with the goal of having one edge per snapshot, where the original presence bit was set to 1. At the end, the total number of resulting edges was about 42G edges.

Dblp[8]**.** This is the well known bibliographic database that keeps track of all the publications in Computer Science including conferences, journals and books. We built the co-authorship network using only papers in conferences and journals, using the years as different dimensions for the network.

We ended with about 600k nodes and 3M edges over 29 dimensions, corresponding to years 1979-2007.

Table **??** summarizes the main characteristics of the network used in the next section. All the statistics are calculated on the aggregated networks, i.e. on the network considering all the defined dimensions. In Column 1 we find the name of the dataset; Column 2 specifies the entities used as nodes in the network; Column 3 roughly summarizes the defined dimensions (please refer to the above list for a detailed description of the dimensions); in Column 4 we find the total number of dimensions in the network; Column 5 and 6 show the total number of nodes and edges for each network; Column 7 shows the average degree (edges/nodes), while Column 8 indicates the "density" of the network, expressed in number of edges over the number of possible edges, taking into account also the number of dimensions.

We can see how low density values we have, but this is the effect of computing this metric taking into account the number of dimensions. The aggregate also makes possible to have very high values of average degrees.

5.2 Tools

All the statistics were computed on two different machines: a server with 4 Intel Xeon processors at 2GHz, equipped with 16GB of RAM , running GNU/Linux 2.6.27 for the experiments on the WebUk and Flickr

[7] http://law.dsi.unimi.it

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

datasets, and a laptop equipped with a Intel Core2 Duo processor at 2GHz with 3GB of RAM, running GNU/Linux 2.6.28, for all the other datasets. All the statistics were implemented in Java, making use of the WebGraph[9] and fastutil[10] libraries. The running times were about 1.5 hours for the Flickr and WebUk datasets, and less than 5 minutes (usually a few seconds) for all the others. The memory occupation was between 6 and 12 GB for Flickr and WebUk, respectively, and less than 1GB for all the others.

6 Experiments

In this section we present some experiments performed after modeling several real-world networks with multigraphs. We show how to apply the metrics we have defined in Section **??** to data coming from very heterogeneous sources: exchange of emails, online social networks, the Web, query log of a search engine, digital bibliographies, etc. Our aim is not to perform an extensive analysis of each of these datasets, but rather to show how powerful and meaningful the metrics, or a combination of them, are in several different contexts. We explicitly put the emphasis on the new metrics we have defined, to show the power of a true multidimensional analysis, that take several kinds of relations into account at once. In the future we plan to analyze some of these datasets under different points of view, trying to extract, thanks to our measures, new laws and models.

6.1 Multidimensional Degree

We start our experimental section by applying the multidimensional degree measure, as defined in Definition **??**, on the two email datasets: Enron and Newsgroups. Viewed from different perspectives, they can be either very similar or very different. For sure, as they both track the exchange of information among users through the same medium, the email, they appear to be similar in their essence. However, their contexts differ: while Enron is a collection taken from the context of a company, so a group of people working together, the kind of entourage of the people involved in the Newsgroup dataset is much different, as the users write about their hobbies, their habits, their religious or political thoughts, and so on, and this is most likely to be done in the free time.

Figures **??**a and **??**a actually confirm exactly this diversity: in these plots we see the dimension degree for single dimensions (days of the week) in the two datasets. As we can see, while people that work for Enron tend to send emails only during weekdays, Newsgroups are populated by messages that arrive every day, without a significant distinction among days.

This expected result opens the way for a possible application of our analysis, namely compressing or sampling a complex network, while keeping specific global properties such as the degree. In fact, from the figures, it is possible to identify different "clusters" of degree distributions: for the Enron dataset this is done by separating "Saturday" from "Sunday" and "rest of the week". After this step, picking up only good representative for each cluster (specific nodes or entire dimensions) would translate in a focused sampling, or would allow for a (lossy) compression of the graph, which are both interesting problems nowadays [**?**,**?**]. Clearly, this process is more effective when the clusters are very clear, assumption that for example does not hold for Newsgroups, where one single dimension is a good representative of the entire network.

If we look at the cumulative dimension degree distribution (Figure **??**b and Figure **??**b), we may note another effect, related to the above one. The figures report the cumulative degree distributions starting from "Saturday". It can be noticed that in Enron considering only the first three dimensions at once produce a cumulative distribution very close to the one obtained taking the complete set of dimensions, while, again, this effect is not appreciable in Newsgroups. Again, this effect help in finding *representative* dimensions for the whole network, w.r.t the degree.

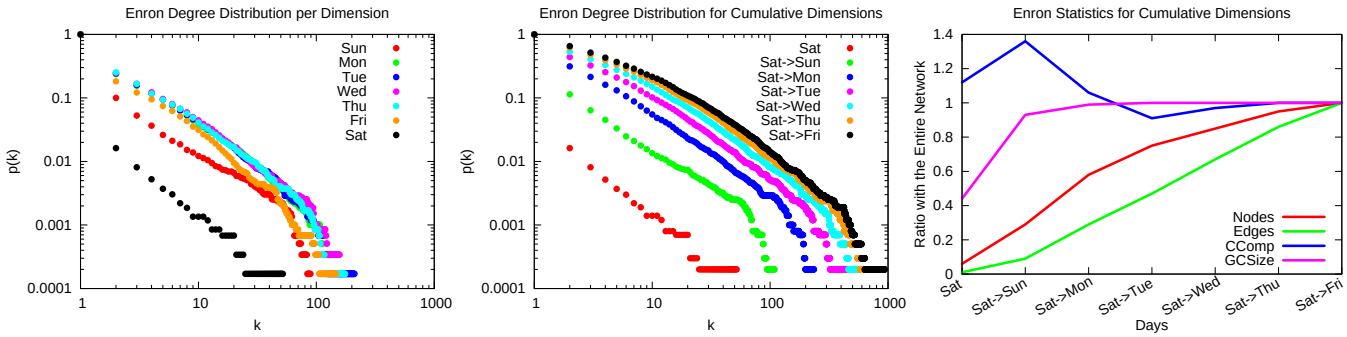
6.2 Connected Components

In the same optic of capturing the differences between the flow of information in Enron and Newsgroups we analyzed some basic statistics including the number of Connected Components of the networks. In Figure **??**c and Figure **??**c, it is possible to see the values of some global properties of the networks, starting by considering only the dimension "Saturday", then adding one by one all the other days, using the function defined in Definition **??**.

We have depicted: the ratio between the number of connected components in each aggregate and the number of connected components in the entire final network; the ratio between the number of nodes (or the number of edge) in the aggregate and the number of nodes (or the number of edges) in the entire final network; the ratio between the size (in terms of nodes) and the giant component of an aggregate and the size of the aggregate itself.

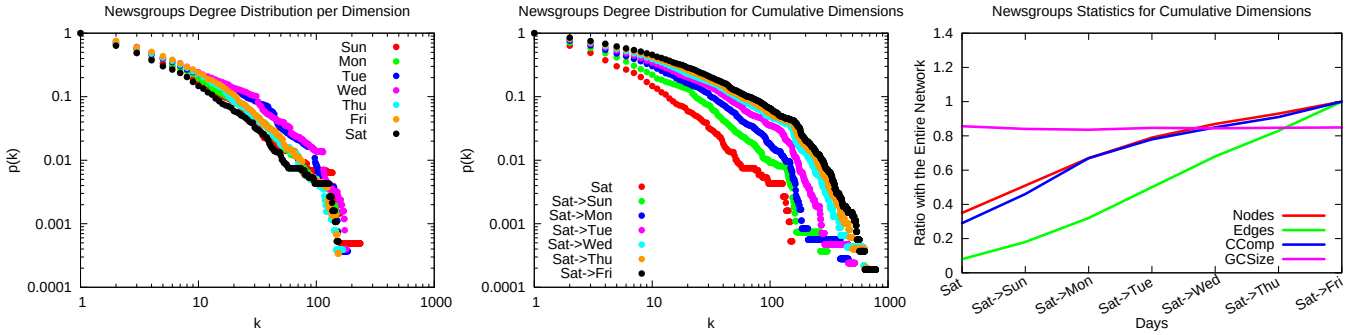
It can be noticed that while in both datasets the number of edges increases in a very similar way, the

[9] http://webgraph.dsi.unimi.it
[10] http://fastutil.dsi.unimi.it



(a) (b) (c)

Fig. 3 The cumulative degree distribution (a), the degree distributions for cumulative dimensions (b) and some basic statistics of the network for cumulative dimensions (c) for the Enron dataset (color image).



(a) (b) (c)

Fig. 4 The cumulative degree distribution (a), the degree distributions for cumulative dimensions (b) and some basic statistics of the network for cumulative dimensions (c) for the Newsgroup dataset (color image).

number of nodes follows a slightly different behavior. We can see, in fact, that the number of nodes in Enron increases very fast in the weekend (60% in three days), while the growth is slower during the rest of the week (remaining 40% in four days). On the other hand, in Newsgroups, the growth is more linear during the whole week.

The two datasets are different also in terms of number of connected components and size of the giant component (which are strongly correlated to each other): while in Enron we have to wait until Monday for having a dense network that grows (almost) linearly, this is not the case in the Newsgroups, where the giant component represents always the same percentage of the aggregate, and the number of connected components grows in a linear trend.

6.3 Neighbors

The number of neighbors is a measure closely related to the degree of a node (see Definition **??**). We can say that in multidimensional networks it plays the same role as the degree for the monodimensional setting. It is therefore natural to expect that it follows a power law in most of the networks, as the degree does. We analyzed this measure on two networks: Figure **??**a shows the results computing the metric on DBLP. It is noticeable that the degree computed on the aggregate of all the dimensions follows a power law. In this plot, moreover, the number of neighbors and the average degree on dimensions (Definition **??**) follow the same distribution. In all the data we tested, we found the distribution of the degree very related to the one of the number of neighbors: when the first was a power law, so was the second, and viceversa. This is easy to explain, recalling Theorem **??**.

We tested this measure also on Flickr. Being a particular kind of online social network, we were expecting for power laws. Surprisingly, as we can see in Figure **??**b, the aggregate degree does not seem to follow such a distribution. The number of neighbors behaves, as expected, following the same distribution of the degree. The explanation can be found by observing the plot in Figure **??**c that represents the degree distributions of the four dimensions of the network. It is possible to see how the degree distribution of the aggregate is very similar to the dimensions “Comments” and “Favorite”. These two dimensions, in fact, cover a very large ma-

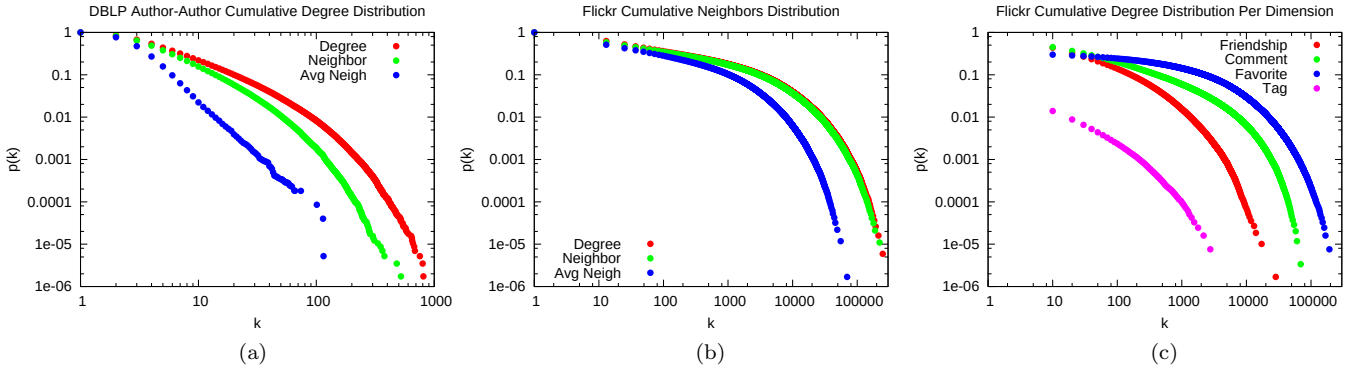


Fig. 5 The cumulative degree, neighbors and average neighbor distributions for DBLP (a) and Flickr (b) datasets, and the degree distributions per dimension for Flickr dataset (c) (color image).

jority of edges in the network, and the proportions with the other two dimensions are impressive: "Favorite" has roughly 675M of edges, "Comments" about 200M, while "Friendship" has only 45M edges and "Tag" 2M. The dimension "Friendship", taken alone, shows, as expected in social contact type of relations, a power law. We can conclude that in a multidimensional network it may exist one or more dimensions that show a scale free degree distribution even if the whole aggregated network follows a very different behavior.

6.4 Dimension Relevance

The dimension relevance is a measure that helps in understanding at the micro level of a single node which dimensions are important for the connectivity of that node to the network (Definitions **??**, **??** and **??**). It is, however, possible to infer global information about the connectivity of the whole network by looking at the cumulative distribution of dimension relevance.

In Figure **??** we report the cumulative distributions of the three different variants of the dimension relevance computed on the Flickr dataset. We recall that there is a sensible difference among the three variants, thus we expect them to produce different values for each node and dimension.

However, it is possible to see that the dimension "Comment" follows the same distribution for each of the three measures, and so does also the dimension "Tag". Recalling that the dimension relevance XOR is always lower than the other two variants (Theorem **??**), this means that, no matter if the dimension is important ("Comment") for a node or not ("Tag"), given that the three distributions are similar the pairs that are connected in those dimensions are not connected in the others.

The remaining two dimensions show a different behavior. Figures **??**a, **??**b and **??**c show that the distribution followed by "Friendship" for the weighted dimension relevance is always under the one followed for the dimension relevance, and that the distribution of the dimension relevance XOR stays below the two. We noted that in Flickr the "Friendship" dimension is not likely to have high values of the dimension relevance XOR, which means that, usually, if two users are friends, they have also tagged, commented, or set the same picture as favorite.

Similar considerations can be done also for "Favorite": it seems to be common for users to share the same picture as favorite with other users, but it is rare that this will be the only kind of relation between any two users. In addition, while, as we said above, "Comment" and "Tag" keep the same distribution when considering the XOR variant, both "Comments" and "Favorite" show lower curves, but "Favorite" tends to go to zero more quickly.

6.5 Totally Split and Totally Mixed

As one can see from Definition **??** and **??**, the concept of Totally Mixed and Totally Split is directly derived from the combination of multidimensional Degree and Neighbors. We recall that a node is said to be totally split if its degree is equivalent to the number of its neighbors, while it is said to be totally mixed if its degree is equal to the number of its neighbors multiplied by the number of the dimensions of the network.

Being totally mixed or totally split for a node has different meanings depending on the semantic of the dimensions in the network. In Figure **??** we have represented the neighbors of three nodes, a totally split (Figure **??**a) and two totally mixed (Figure **??**b and **??**c), found in the network obtained from the Query log dataset.

In this network a totally split node is a word that is linked to all its neighbors by only a specific rank: it hence allows to identify the exact words with which that word always produces good results (high rank) and

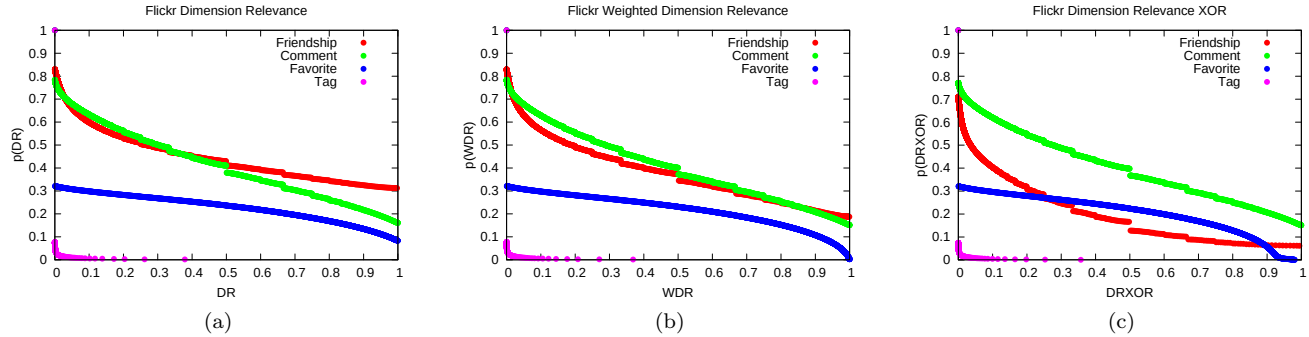


Fig. 6 Dimension Relevances for Flickr dataset (color image).

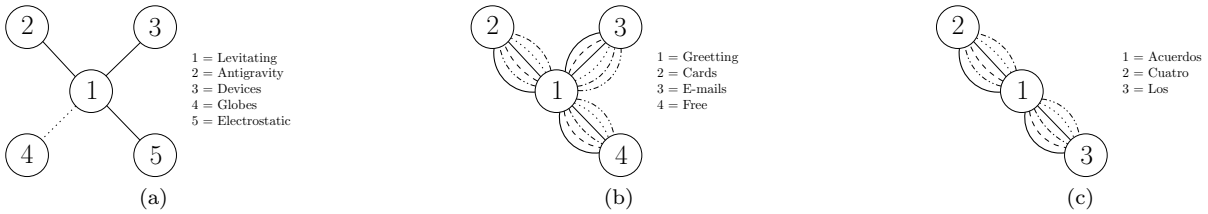


Fig. 7 Examples of totally split and mixed (from misspelt and foreign words) nodes found in the Query Log dataset.

those with which it always works bad (rank very low). Figure **??**a shows a clear example of this situation: the word "Levitating" is found to work well when used in conjunction with "Devices", "Antigravity" and "Electrostatic", but it seems to work bad when used together with "Globes".

On the other hand a totally mixed node means that, regardless of which other words were used together in a query, the users felt the need to check all the results provided by the search engine. Words of this kind were verified being either typing errors (Figure **??**b: "Greetting") or non-English words (Figure **??**c "acuerdos"). We believe that this approach may significantly help the world of the search engines in applications such query recommendation or refinement [**?**], clustering of the search results [**?**], or simply improving the quality of the search results.

6.6 Clustering Heterogeneity

Another interesting side of the analysis of the interactions among different dimensions is given by Clustering Heterogeneity. Using this metric it is possible to understand as different dimensions contribute when forming triangles in the network (see Definitions **??** and **??**). The higher this value the more "important" is the role of a single dimension in building a multidimensional triangle (defined in Definition **??**), w.r.t the fact of being the only dimension connecting two nodes in the triangle itself.

We tested this metric on two datasets: Enron and Newsgroups. For the second dataset we considered two separated networks, one in which the dimensions represent the days of the week and the other one in which they represent the newsgroups. Looking at Table 4 it is possible to make a few considerations: in column 1 we have the dataset, in column 2 the dimensions considered, column 3 shows the number of multidimensional triangles found, while column 4 shows the value of the cluster heterogeneity. The first consideration is that both Enron and Newsgroups in these temporal dimensions behave in a similar fashion. Multidimensional triangles in Enron are still very rare and/or take place amongs the same people, while in Newsgroups are much more common. The similarity comes with the Clustering Heterogeneity values: in Enron just over half and in Newsgroups almost three quarters of the edges of these triangles are created on a particular day.

Dataset	Dim Set	MultiDimΔ	CH
Enron	Days	6,075	0.58
NewsG	Days	754,910	0.74
NewsG	Newsgroups	1,872	1

Table 4 Cluster Heterogeneity in various datasets.

If we compare the two Newsgroup networks, one with the days, one with the newsgroups, we can highlight a particular behavior: the number of closed multidimensional triangles in the second network is much smaller, i.e. Newsgroups is a network in which people that answer together to the same post in a newsgroup are not likely to answer together to a post in a different newsgroup. This is also highlighted by the value of the cluster heterogeneity: in the few multidimensional triangles, there are no pairs of people connected by

more than one newsgroup. Please note that this does not mean that who writes in a newsgroup, does not write in another one: in this case there would not exist multidimensional triangles. Instead, the stress is in the *connection* between two people: that is found to belong only to one dimension.

6.7 Dimension Degrees

The analysis of the dimension degree (Definitions **??** and **??**) is helpful to understand which dimensions contain which percentage of the edges in a network. Besides, the dimension degree uniqueness (Definitions **??** and **??**) tells how many of the edges of a dimension are the sole link between two nodes, enriching the meaning of the previous measures.

These measures can be used, also in conjunction, with several different aims. In particular, our intent is to show how they can be used to perform basic analysis of the temporal evolution of a network, thus making the dynamic temporal analysis of a network a particular case of multidimensional analysis. In order to do so, we need that each dimension of the network expresses a temporal snapshot of the network itself. Hence, we conducted this analysis for the WebUK and DBLP dataset, the first having twelve monthly dimensions, the second having 29 yearly dimensions from year 1979 to 2007.

Figures **??**a and **??**b show the different behaviors of WebUK and DBLP. While DBLP presents a constant growth in the number of edges, the growing behavior WebUK is more irregular. Although this may be partly affected by the problems encountered by the crawler that produced the dataset (documented in [**?**,**?**]), the general behavior in this dataset is different from DBLP.

However, it is important to note the different time span of the two datasets: while we have data for 29 years in DBLP, the monthly snapshots of WebUk take into account only one year. From the statistics, summarized in the table in Figure **??**c, we know that the ratio, both in terms of nodes and in terms of edges, between one year of DBLP (we consider the 2006) and the previous one is lower than the ratio between the end of the WebUk year and its beginning, i.e. WebUk grows faster than DBLP.

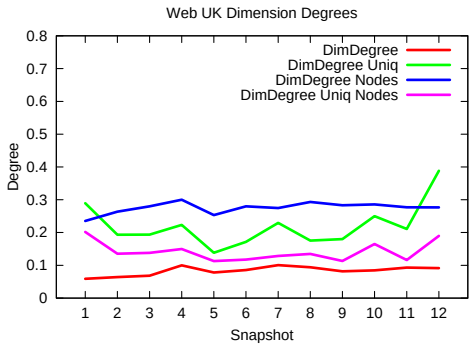
In WebUK the dimension degree uniqueness represents the portion of the link in a month that does not appear in any other month. In this case the values are heavily accentuated by the problems in the crawling. Please note that the first and the last snapshot behave in a different ways and appear to have higher vaules, because their edges or nodes are not present in the past (for the first snapshot) or in the future (for the last). It is interesting noting that there are some peaks in Figure **??**a. These peaks represent snapshots of the network (4, 7 and 10, i.e. September, December and March), where a greater portion of the edges are no longer found in the network, effectively making their contributions less important to the final topology of the network. We might think about those as temporary trends of the Web, maybe they are links or pages following some important news that was discovered to be false later.

Other considerations are possible on the DBLP dataset. In Figure **??**b it is possible to see that the dimension degree uniqueness follows an irregular decrease from 1979 to 2000, while starting 2001 it increases again. Representing the number of authors who published only in a particular year (nodes), or the number of collaborations that took place only in a specific year (edges), the dimension degree uniqueness tells us that, starting from year 2001, the occasional scientific contribution of authors from outside the DBLP community has been playing an important role in the global network.

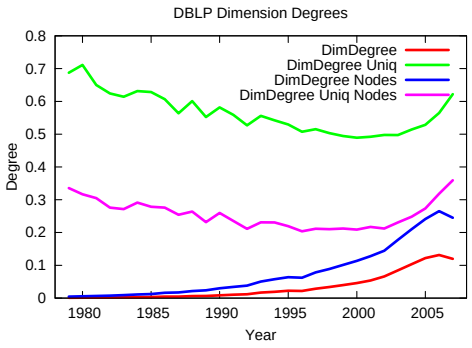
6.8 Dimension Correlation

We tested this metric both on the Supermarket dataset and on DBLP, where we considered both the node (Definition **??**) and the edge (Definition **??**) variants of this measure. While, however, the node variant gives a view of the correlation of two dimenions considering the actors of a network, the edge variant puts more emphasis on the relationship among them. Being interested in the analysis of groups of interacting people, we show here the results obtained with the edge variant. The other variant, however, gave pretty similar results.

Figure **??**a and Figure **??**b represent very simplified (i.e., they consider only a few representative dimensions of the networks) graphs of the relationships found among some of the dimensions inferred from the values of the correlation. As we can see, in Figure **??**a it is possible to identify *clusters* of interacting dimensions: “Self Service” (i.e., in a supermarket, the department where you can pick up by yourself the desired quantity of fruit and vegetables) and “Fresh” , “Bread”, “Fruit” (that includes vegetables) and “Very Fresh” (i.e., food prepared daily by the supermarket itself), “Container” and “Packaged”. Inside each cluster, we see very high values of the correlations among the dimensions (in one cluster this value is 1), while the correlations between clusters (the double arrows in this graph) are very poor. The high values are due to the implicit semantic of the dimensions (that is taken from the metadata of the original database): bread, fruit, and very fresh are similar to each other, while bread and container do not have (almost) anything to do with each other. In Figure **??**c are represented the values of edge correlations between the dimension used for the Figure **??**a, that confirm



(a)

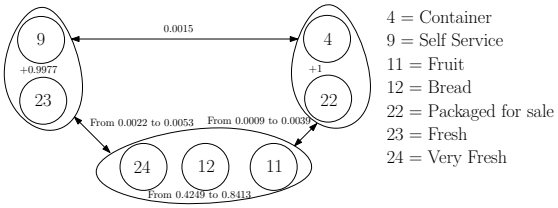


(b)

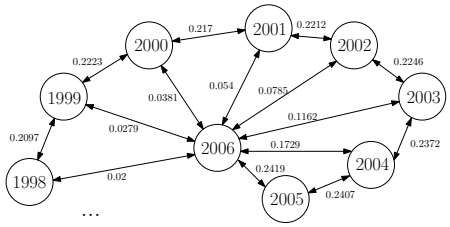
Dataset & Dim.	Nodes	Edges
DBLP 2006 / DBLP 2005	1.09	1.07
WebUk May07 / WebUk Jun06	1.17	1.54

(c)

Fig. 8 Dimension degrees for WebUk and DBLP datasets (color image).



(a)



(b)

Fig. 9 The correlation graph of the Supermarket and DBLP datasets.

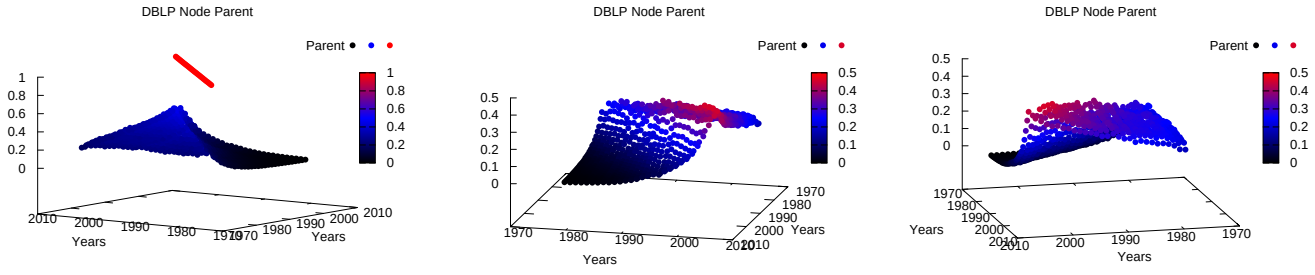
our analysis. Figure **??**b shows the same kind of graph in DBLP. In this case the situation is pretty different: there are no cluster, even though there is another effect, i.e. there is a relatively high correlation between consecutive years, while it goes to zero when taking two years far from each other. Another interesting point is that, generally speaking, the correlation between pairs of consecutive years tend to increase in more recent pairs: this is due to the fact that the DBLP network generally increases.

Please note that in this kind of analysis, we implicitly give the basis for another interesting research direction: as a cluster can be, in fact, replaced by a metanode representing it, we are actually performing the detection of *hierarchies* in networks, sometimes called *multilevel* analysis [**?**], graph grammar extraction [**?**], and other similar views of the same problem. An important difference, however, with the graph OLAP, is that we perform *semantic* clustering, i.e. we are able to cluster also dimensions that have no particular ordering among them. In order to aggregate them, instead of relying of some particular property, such as "2001" is close to "2002", we base our clusters on the extracted values of the metrics, (almost) without apriori knowledge of the meaning of the dimensions.

6.9 Parent

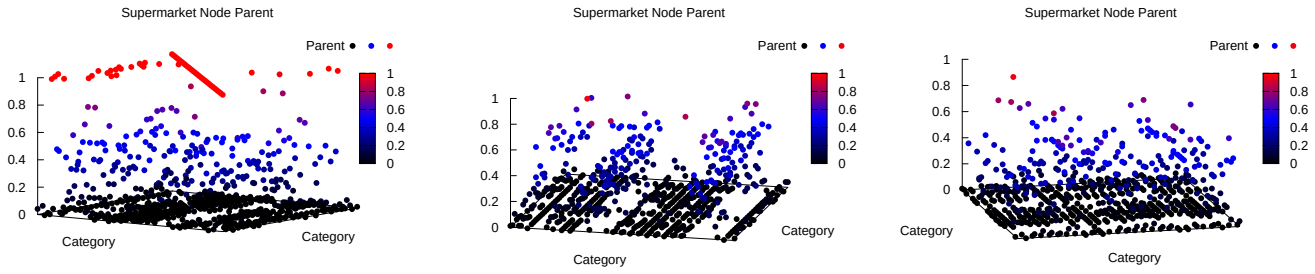
This measures reveal situations in which a dimension "includes" another one (in terms of nodes, Definition **??**, or in terms of edges, Definition **??**). However, rather than defining it as a boolean function, we preferred to let it take all the possible values between 0 and 1, in order to detect this phenomenon also at intermediate levels.

Consider the Figure **??** and Figure **??**, where each of them shows three views of the same plot. In in these figures we represent the values of Node Parent computed on two different datasets: DBLP and Supermarket. In DBLP, the straightforward hypothesis is that each year includes almost all the authors of the previous year, excluding a few people who have stopped publishing, and adds some newcomers. We obtained proof of this in our results, where, for example, it is noticeable that the value of the parent between 2007 and 2006 is very high and clearly higher than the one between 2007 and 1987. Being asymmetrical, the parent between 2006 and 2007 is not equal to the one between 2007 and 2006, and actually we found it lower. The explanation can be found if we consider that there are more new authors than the ones who stop publishing papers. The plot in Figure **??**a and **??**b shows then two different slopes in its two sides, while having the bisector of the plane increasing towards the most recent years: the 2007 include 2006 more than how 2001 includes 2000.



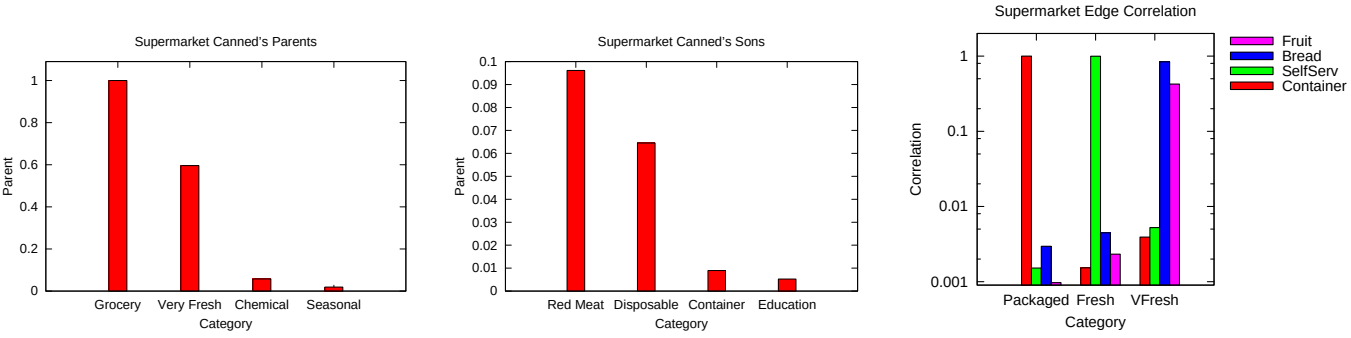
(a) (b) (c)

Fig. 10 Node parent on DBLP dataset. (b) and (c) plots do not include the parent of a dimension applied to itself, always equal to 1 (color image).



(a) (b) (c)

Fig. 11 Node parent on the Supermarket dataset. (b) and (c) plots do not include the parent of a dimension applied to itself, always equal to 1. Please notice that X and Y axis have no particular order, since they represent a marketing category of the retail company (color image).



(a) (b) (c)

Fig. 12 Examples of parent and correlation on the Supermarket dataset (color image).

While the values of the parent for the DBLP dataset depict a scenario easy to understand, this is not the case for the Supermarket dataset. Please note that the plots in Figures **??**a and **??**b show values of the parent for a set of dimensions in which is not possible to find a reasonable order: while in DBLP two years are clearly comparable, defining an order between "Bread" and "Container" would be not meaningful. Hence, a "cloud" shape of values depicted by the parent is expectable. Nevertheless, while in DBLP there is no explicit hierarchy among the dimensions, this is not true for Supermarket, while the products are arranged explicitly into categories, departments, and so on. In this case, in fact, it is easy to validate the parent analysis: we expect its value to be equal to one in case of an explicit hierarchical inclusion. Moreover, given the definition of this metric, any value between zero and one expresses the "parent" relationship at different levels. Figure **??** shows some "parent" and "child" relations involving the dimension "Canned". In Figure **??**a we have represented the two dimensions with the highest parent value for "Canned" (left), and the two with

the lowest (right). As expected, "Grocery" is a clear parent of "Canned", but we were also able to find another dimension with which a slightly poorer parent relation holds: "Very Fresh". This means that the customers who buy food in cans are also included in those who buy very fresh food at the supermarket, like bread. Two dimensions which are completely separated from "Canned" are "Chemistry" and "Seasonal": who buys canned food is not looking for seasonal food.

Being "Canned" not a macro-category for goods, its parent values, when detecting its children, are lower than the ones found for "Grocery", but still in Figure **??**b, is possible to see how this dimension includes some others: "Red meat" and "Use disposable". Totally different dimensions are "Containers" and "Education and Entertainment".

7 Conclusions and future work

In this paper, we considered the problem of analyzing *multidimensional* networks. After characterizing such networks by means of *multigraphs*, we systematically defined the main analytical metrics. We have both extended to the multidimensional case well known metrics, broadly used in social/complex network analysis and graph theory, and introduced brand new ones, which exploit explictly the multiple dimensionality – and therefore only make sense in the multidimensional case. We have demonstrated the analytical power of the new metrics, whose main feature is to capture the interplay among different dimensions of the same network. 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 metrics; as demonstrated in this paper, several interesting analytical questions, which need to investigate relationships between network dimensions, can be answered in a natural way by the proposed mechanisms.

On the other hand, we are aware that the research described in this paper leaves many problems open for futher research, both on the theoretical and the application side. Is the repertoire of metrics sufficiently wide to express the desired class of analytical questions? Are there interesting properties of the metrics that may help the analysis, or be exploited for optimizing the computation of the metrics 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 We would like to thank Alessio Orlandi for helping us managing the WebUk dataset, and for the help with the fastutil library. We are also grateful to Claudio Lucchese who gave us part of the Flickr data. Moreover, we would like to thank all the persons that made the data we used in this paper publicly available: thank you very much, you help making the global research possible and interesting with your efforts!