# The Pursuit of Hubbiness: Analysis of Hubs in Large Multidimensional Networks

Michele Berlingerio[a], Michele Coscia[a,b], Fosca Giannotti[a], Anna Monreale[a,b], Dino Pedreschi[b]

[a] *ISTI - CNR, Area della Ricerca di Pisa, Italy - {name.surname}@isti.cnr.it*
[b] *Computer Science Dep., University of Pisa, Italy - {coscia,annam,pedre}@di.unipi.it*

**Abstract**

Hubs are highly connected nodes within a network. In complex network analysis, hubs have been widely studied, and are at the basis of many tasks, such as web search and epidemic outbreak detection. In reality, networks are often multidimensional, i.e., there can exist multiple connections between any pair of nodes.In this setting, the concept of hub depends on the multiple dimensions of the network, whose interplay becomes crucial for the connectedness of a node. In this paper, we characterize multidimensional hubs. We consider the multidimensional generalization of the degree and introduce a new class of measures, that we call Dimension Relevance, aimed at analyzing the importance of different dimensions for the hubbiness of a node. We assess the meaningfulness of our measures by comparing them on real networks and null models, then we study the interplay among dimensions and their effect on node connectivity. Our findings show that: (i) multidimensional hubs do exist and their characterization yields interesting insights, and (ii) it is possible to detect the most influential dimensions that cause the different hub behaviors. We demonstrate the usefulness of multidimensional analysis in three real world domains: detection of ambiguous query terms in a word-word query log network, outlier detection in a social network, and temporal analysis of behaviors in a co-authorship network.

*Keywords:* Complex Network Analysis, Social Networks, Graph Theory, Hubs, Multidimensional Data.

## 1. Introduction

Complex networks have been receiving increasing attention by the scientific community. One reason for this is the availability of massive network data from diverse domains, and the outbreak of innovative analytical paradigms, which pose relations and links among entities, or people, at the center of investigation [12, 13, 21, 1, 27, 5, 2, 28]. One topic of research in this direction has received considerable attention from the scientific community: finding and analyzing *hubs*, i.e., nodes with a large number of neighboring nodes.

Most of the networks studied so far are monodimensional: there is only one interaction among nodes. In this setting the concept of hub has been widely studied, and is at the basis of many important applications, ranging from analysis of the structure of the Internet to web searches, from peer-to-peer network analysis to social networks, from Viral Marketing to analysis of the Blogosphere, from outbreaks of epidemics to metabolic network analysis [4, 14, 1, 13, 11, 24, 15, 17].

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

Dimensions in network data can be either *explicit* or *implicit*. In the first case the dimensions directly reflect the various interactions in reality; in the second case, the dimensions are defined by the analyst to reflect different interesting qualities of the interactions, that can be inferred from the available data. This is exactly the distinction studied in [18], where the authors deal with the problem of community discovery. In their paper, our conception of multidimensional network is referred as *multislice*, networks with explicit dimensions are named *multiplex*, and the temporal information is used to derive dimensions for the network.

Examples of networks with explicit dimensions are social networks where interactions represent communications by different means: email, instant messaging services and so on. An example of network with implicit dimensions is a co-authorship network where an interaction between two authors represents the time when the collaboration took place.

In this paper, we deal with the following question: *how does the concept of hub change in multidimensional network analysis?* Figure 1 depicts a possible hub in a monodimensional network (Figure 1a) and three possible hubs in a multidimensional setting (Figure 1b-d). The four cases show different hub configurations: while the first is

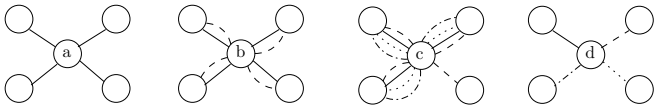



Figure 1: Example of different multidimensional hubs.

simply a node with a high degree (thus connectivity), and nothing else can really be said about it purely on the basis of this figure, the other three represent a different scenario. We can see that the hub in Figure 1b is connected by two dimensions (solid line) to all the other nodes, while this is not true for the other hubs. Neither the degree of the node nor the number of neighbors that could be reached from it would give us any more information. The third and fourth case give other two possible scenarios, where, if we take into account each dimension individually, the node in the center has a low degree (and number of neighbors); however, the co-existence of many dimensions where this happens makes it possible to consider the central node as a hub (this is particularly true for the hub in Figure 1d).

Can the four hubs be considered in the same way, or can we say something specific about each one? In a multidimensional setting, are all hubs equivalent to each other? Can we say something about the importance of a specific dimension for the connectivity of a node? Finally, can we reason on hubs' behavior by looking at how relevant a dimension is for the connectivity of the hubs?

As these questions suggest, analyzing hubs in multidimensional networks basically introduces a new degree of freedom: the set of dimensions of the network. However, we believe that the current analytical tools are not able to capture the interplay among these dimensions. New measures need to be introduced to overcome this problem.

In this paper, we address the problem of finding and analyzing multidimensional hubs in real networks by defining suitable analytical tools. This work is only a part of our research work on multidimensional network analysis, for which we posed the bases in a previous work [6], by defining a new model and new measures for multidimensional networks.

The contribution of this paper can be summarized as follows. First, we introduce multidimensional networks and we show some real world examples of them. Next, we define the problem of finding and analyzing multidimensional hubs in such networks. Further, we introduce two analytical tools needed in order to perform such an analysis. The first is a multidimensional generalization of the degree, namely the number of neighbors of a node, while the second is a brand new class of measures, which we call *Dimension Relevance*. The aim of these measures is to exploit the additional degree of freedom that multidimensionality adds to the problem of analyzing hubs in networks. Finally, we show a multidimensional hub analysis case study on the proposed real world networks, supporting the meaningfulness of the problem introduced, the effectiveness of the measures defined, and a few practical applications intended to demonstrate the power of our approach.

The most important results of this work are: (1) we show that multidimensional hubs exist, and can be found and analyzed using our introduced measures of interplay of the different dimensions; (2) we show that the characterization of multidimensional hubs highlights interesting analytical properties, and (3) thanks to our metrics, we discover and quantify the importance of every single dimension with respect to the others, generally unknown a priori.

The remainder of the paper is organized as follows: Section 2 introduces the multidimensional network setting and gives a formal definition of the problem under investigation; Section 3 introduces the various measures for multidimensional hub analysis and a possible characterization of multidimensional hubs; in Section 4 we present the results about the analysis of the proposed measures by applying them on real networks; Section 5 presents three examples of characterization of hubs in real networks; Section 6 overviews previous related work; finally in Section 7 we conclude our work with recommendations for possible future research.

## 2. Multidimensional Networks in Reality

The term *network* denotes a structure that is made up of a set of entities and connections among them. A network with connections of different kinds is called a *multidimensional network*; we use the term *dimension*, instead of link type or kind, to emphasize that each link dimension corresponds to a different perspective of the network connectivity structure.

Most real life networks are intrinsically multidimensional, and some of their properties may be lost if the different dimensions are not taken into account [3]. In other cases, it is natural derive multiple link dimensions from the available data to the end of analyzing some phenomena.

### *2.1. Three Real-World Examples*

Three examples of real-world multidimensional networks, highly heterogeneous and representative of the possible different kinds of networks in the real world, which we acquired and prepared as the subject of our study, are the following:

**Flickr**[1]**.** This dataset comes from the well known photo sharing service, and was obtained by crawling the data via the available APIs. We extracted both implicit and explicit dimensions of the social network represented in this data. For each picture, we extracted the list of all the users related to it and from these users we completed the social network by adding edges if two users commented, tagged or

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

set the same picture as favorite, or if they had each other as a contact. From roughly 1.3M users we obtained slightly more than 900M edges, distributed on the above mentioned four dimensions. The resulting network is a person-person network, where each dimension is one of the "Friendship", "Tag", "Favorites", or "Comment", representing if the users are friends, tagged the same picture, marked the same picture as favorite, or commented on the same picture. A small extract of this network is represented in Figure 2(a).

**DBLP[2].** This dataset comes from the popular bibliographic database. We constructed a co-authorship network of authors (nodes) connected by an edge if they wrote a paper together. We used years as dimensions, and any pair of authors was connected in a specific dimension if they wrote at least one paper together in the corresponding year. We obtained roughly 600k nodes connected by 2.6M edges, distributed over 65 dimensions. The resulting network is a person-person network, where each dimension is on the years from 1938 to 2008 (with some gaps at the beginning), indicating whether the users had a collaboration in the corresponding year. A small extract of this network is represented in Figure 2(b).

**Query Log[3].** This network was constructed from a query-log of approximately 20 millions web-search queries submitted by 650,000 users over a period of time, and was described in [20]. Each record of this dataset stores an anonymous user ID, the query terms, the date and hour of the query, the rank position of the result visited by the user on each record and the host portion of the URL of the visited result. From this dataset, we extracted a word-word network of query terms, consisting of roughly 200k words (nodes), after removing stop-words. We connected two words if they appeared together in a query, producing roughly 2M edges. Dimensions are defined as the rank positions of the results, grouped into six almost equi-populated bins: "Bin1" for rank 1, "Bin2" for ranks 2-3, "Bin3" for ranks 4-6, "Bin4" for ranks 7-10, "Bin5" for ranks 11-58, "Bin6" for ranks 59-500. Hence two words appeared together in a query for which the user clicked on a resulting url ranked #4 will produce a link in dimension "Bin3" between the two words. The result is a word-word network, for which we give a small extract in Figure 2(c).

Table 1 summarizes the main properties of these networks (see caption). As we see, while in Flickr the dimensions are explicit, in QueryLog and DBLP we have to define our concept of dimension, thus in this case the dimensions are implicit.

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

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

| Dataset | Dimension | #Nodes | #Edges | $k$ | Density |
|---|---|---|---|---|---|
| QueryLog | Bin 1 | 138,992 | 1,104,581 | 15.894 | $1.14e^{-4}$ |
| | Bin 2 | 108,439 | 878,136 | 16.195 | $1.49e^{-4}$ |
| | Bin 3 | 89,418 | 708,897 | 15.855 | $1.77e^{-4}$ |
| | Bin 4 | 75,846 | 583,774 | 15.393 | $2.02e^{-4}$ |
| | Bin 5 | 42,951 | 253,976 | 11.826 | $2.75e^{-4}$ |
| | Bin 6 | 12,236 | 36,456 | 5.958 | $4.87e^{-4}$ |
| | **Global** | **184,760** | **3,565,820** | **38.599** | $3.48e^{-5}$ |
| Flickr | Friendship | 984,919 | 48,723,010 | 98.938 | $1.00e^{-4}$ |
| | Comment | 930,526 | 198,309,709 | 426.231 | $4.58e^{-4}$ |
| | Favorite | 380,992 | 674,488,956 | 3540.698 | $9.29e^{-3}$ |
| | Tag | 91,690 | 715,447 | 15.605 | $1.70e^{-4}$ |
| | **Global** | **1,186,895** | **922,237,122** | **1554.033** | $3.27e^{-4}$ |
| DBLP | **Global** | **582,201** | **2,648,845** | **9.09** | $7.81e^{-6}$ |

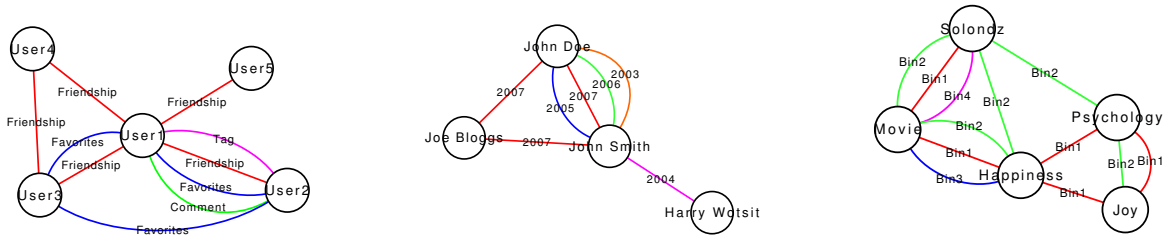
Table 1: Summary of the datasets used. Column 1 specifies the dataset; Column 2 the dimension into account; Columns 3 and 4 the number of nodes and edges; Column 5 the average degree; Column 6 the density computed as number of edges out of number of total possible edges in all the dimensions

### *2.2. Finding and Characterizing Hubs*

Most interesting network analytical concepts, both at the global and at the local level, such as connectivity, centrality, diameter, etc., developed for standard, monodimensional networks, come under a different light when seen in the multidimensional setting. At the global level, for example, the connectivity of the whole network changes if we see a single dimension as a separate network, with respect to the network formed by all the edges in the entire set of dimensions. Also at the local level, it is possible to analyze many other examples. One such example is the concept of a hub, i.e., a node with a very high degree, substantially higher than the average degree of all nodes. When considering a multidimensional network, such simple concept becomes subtler and multifaceted: first, the definition of a multidimensional hub is parametric with respect to a set of dimensions and secondly, the relevance of a node depends on the interplay among the different dimensions and their impact on the connectivity of the node. Here, a multidimensional hub is a node with high connectivity in the sub-network obtained by considering only the edges from some specified dimensions (later in the paper we give a formal definition). As evidence of how subtle the characterization of a multidimensional hub is, we found in all our real-world networks that the population of hubs obtained while neglecting the dimensions, differs substantially from that of hubs obtained taking dimensions into account (see Table 2 and its discussion in Section 4.2): some (sometimes many) monodimensional hubs are not multidimensional hubs, and vice versa (see Section 4.2 for further analysis of this phenomenon).

This led us to conclude that analyzing hubs in multidimensional networks is not a trivial extension of the standard case. In other words, it requires techniques and measures of node connectivity across different dimensions, able to highlight the interplay among (sets of) dimensions and their impact on node connectivity. Therefore, the problem that we dealt with in this paper can be defined as follows:

**Definition 1 (Problem Definition).** *Given a large multidimensional network, find and characterize the multidimensional hubs.*




(a) (b) (c)

Figure 2: Small extracts of the three real multidimensional networks.

In the remainder of this paper, we introduce a few analytical measures of node connectivity in multidimensional networks, as a collection of basic tools for hub analytics. We then show in our three case studies how such analytics can be successfully applied to discover sets of multidimensional hubs which highlight interesting phenomena in the associated networks. Our aim is to gain two kinds of insights into the multidimensional network under analysis: (a) the interplay of dimensions and their effect on node connectivity, and (b) the characterization of a group of hubs and the understanding of their role within the network.

Note that, while we are interested in hubbiness, i.e. *degree centrality*, the same kind of approach can be used to analyze the multidimensional versions of other concepts of centrality: betweenness, closeness, and eigenvector centrality are a few examples. We leave for future research the study of those scenarios.

## 3. Connectivity Measures for Multidimensional Networks

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

Since the concept of *hub* is related to the connectivity of a node in the network, to define it properly, we first need to extend the concept of connectivity for multidimensional networks.

Now, we define the *Neighbors* and the *Dimension Relevance* class of measures. Neighbors is an extension of the degree in the multidimensional setting. Dimension Relevance is a new class of measures, meaningful only in multidimensional networks. Moreover, we give their interpretation and we show a toy example illustrating their behavior for a few nodes.

### 3.1. Neighbors

In classical graph theory the *Degree* of a node refers to the connections of a node in a network: it is defined, in fact, as the number of edges adjacent to a node. In a simple graph, each edge is the sole connection to an adjacent node. In multidimensional networks the degree of a node (i.e., the number of the connections of that node in a network) and the number of nodes adjacent to it are no longer related, since there may be more than one edge between any two nodes. For instance, in Figure 1, all nodes have four neighbors, but they have a very different degree, especially in every single dimension.

In order to capture this difference, we define a measure concerning the *neighbors* of a node.

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

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

*where $NeighborSet(v, D) = \{u \in V \mid \exists (u, v, d) \in E \wedge d \in D\}$. This function computes the number of all the nodes directly reachable from node $v$ by edges labeled with dimensions belonging to $D$.* □

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

for the connectivity of that node: removing any of these dimensions would disconnect (directly) that node from some of its neighbors. This is the case of node 5 in Figure 3.

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

**Definition 3 (Neighbors$_{XOR}$).** *Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively. The function $Neighbors_{XOR} : V \times \mathcal{P}(L) \to \mathbb{N}$ is defined as*

$$Neighbors_{XOR}(v, D) =$$

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

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

*3.2. Dimension Relevance*

As already mentioned, while performing hub analysis it is important to understand how important a particular dimension is over the others for the connectivity of a node, i.e. what happens to the connectivity of the node if we remove that dimension. In order to answer these questions, we define the new concept of *Dimension Relevance*.

**Definition 4 (Dimension Relevance).** *Let $v \in V$ and $d \in L$ be a node and a dimensions of a network $G = (V, E, L)$, respectively. The function $DimRelevance : V \times L \to [0, 1]$ is defined as*

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

*and computes the ratio between the neighbors of a node $v$ connected by edges labeled with a specific dimension $d$ and the total number of its neighbors.* □

Clearly, the above function can be defined taking into account a set of dimensions instead of a single dimension. In other words, we can generalize Definition 4 as follows:

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

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

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

Note that, the case of a single dimension (Definition 4) is a particular case of that in Definition 5, where the set of dimensions $D$ contains only the dimension $d$. In the remaining of the paper we define the others measures considering a set of dimensions.

However, in a multidimensional setting, this measure may still not cover important information about the connectivity of a node. Figure 1 shows three nodes ($a$, $b$ and $c$) with a high dimension relevance for the dimension represented by a solid line. In the first two cases the dimension relevance is equal to one, but the complete set of connections they present is different: if we remove the solid line dimension the node $a$ will be completely disconnected while the node $b$ can still reach all its neighbors. To capture these possible different cases we introduce a variant of this metric.

**Definition 6 (Dimension Relevance XOR).** *Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $G = (V, E, L)$, respectively.*
*The function $DimRelevance_{XOR} : V \times \mathcal{P}(L) \to [0, 1]$ defined as*

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

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

We can easily calculate the above metric in the examples in Figure 1. For the node $a$ there is no difference with the *Dimension Relevance* (Definition 5): all its neighbors are only reachable by solid edges. In node $b$ we have the opposite situation: all its neighbors are reachable by solid edges, but we always have an alternative edge. So the *Dimension Relevance XOR* of the solid line dimension is equal to zero.
In the following, we want to capture the intuitive intermediate value, i.e. the number of neighbors reachable through a dimension, taking into account all the possible alternatives.

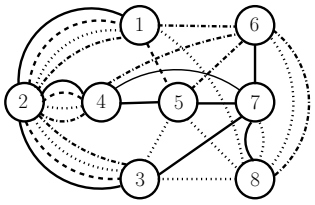
**Definition 7 (Weighted Dimension Relevance).** *Let $v \in V$ and $d \in L$ be a node and a dimension of a network $G = (V, E, L)$, respectively.*
*The function $DimRelevance_W : V \times \mathcal{P}(L) \to [0, 1]$, called* Weighted Dimension Relevance*, is defined as*

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

*where: $n_{uvd}$ is the number of dimensions in which there is an edge between two nodes $u$ and $v$ and that belong to $D$; $n_{uv}$ is the number of dimensions in which there is an edge between two nodes $u$ and $v$.* □

Hereafter we occasionally use DR to stand for Dimensional Relevance. In our toy example in Figure 3, the nodes 6, 7 and 8 have five neighbors, quite a large number in this example, but their values of Dimension Relevance are very different since they are connected in different dimensions.

The Dimension Relevance XOR behaves in a different way. A value equal to zero does not necessary imply that the node is not connected to a particular dimension. It represents a situation where the node has no neighbor that can be reached exclusively through that particular dimension. So it is possible to reach it by alternative ways. In Figure 3, node 3 is an example of this, when considering the dashed line dimension.




| Id | Deg | Neigh | $DR$ | | | | $DR_W$ | | | | $DR_{XOR}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | dim1 | dim2 | dim3 | dim4 | dim1 | dim2 | dim3 | dim4 | dim1 | dim2 | dim3 | dim4 |
| 1 | 7 | 4 | 0.250 | 0.500 | 0.500 | 0.500 | 0.062 | 0.312 | 0.312 | 0.312 | 0.000 | 0.250 | 0.250 | 0.250 |
| 2 | 12 | 3 | 1.000 | 1.000 | 1.000 | 1.000 | 0.250 | 0.250 | 0.250 | 0.250 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3 | 7 | 4 | 0.250 | 0.250 | 0.750 | 0.250 | 0.312 | 0.062 | 0.562 | 0.062 | 0.250 | 0.000 | 0.500 | 0.000 |
| 4 | 7 | 4 | 0.750 | 0.250 | 0.250 | 0.500 | 0.562 | 0.062 | 0.062 | 0.312 | 0.500 | 0.000 | 0.000 | 0.250 |
| 5 | 6 | 6 | 0.333 | 0.166 | 0.333 | 0.166 | 0.333 | 0.166 | 0.333 | 0.166 | 0.333 | 0.166 | 0.333 | 0.166 |
| 6 | 6 | 5 | 0.200 | 0.000 | 0.200 | 0.800 | 0.200 | 0.000 | 0.100 | 0.700 | 0.200 | 0.000 | 0.000 | 0.600 |
| 7 | 6 | 5 | 1.000 | 0.000 | 0.200 | 0.000 | 0.900 | 0.000 | 0.100 | 0.000 | 0.800 | 0.000 | 0.000 | 0.000 |
| 8 | 7 | 5 | 0.200 | 0.000 | 1.000 | 0.200 | 0.100 | 0.000 | 0.800 | 0.100 | 0.000 | 0.000 | 0.600 | 0.000 |

Figure 3: Toy example and computed measures. Lines: solid = dim 1, dashed = dim 2, dotted = dim 3, dash-dotted = dim 4.

The Weighted Dimension Relevance takes into account both the situations modeled by the previous two definitions. Low values of $DimRelevance_W$ for a particular set of dimensions $D$ are typical of nodes that have a large number of alternative dimensions through which they can reach their neighbors. High values, on the other hand, mean that there are fewer alternatives. Our example shows the case of node 4 when considering the solid line dimension: its Weighted Dimension Relevance is clearly the highest, although the dot-dashed line dimension has a high value of Dimension Relevance (as in Definition 5).

The table in Figure 3 shows the values of all the above metrics for all the dimensions computed in the toy example. Each value is computed taking into account a single dimension. In our analysis we will apply our metrics on a single dimension to better highlight and show the use, the effects and the power of proposed measures.

The following theorem states the relations among the above three definitions.

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

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

□

**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. Let:

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

First of all, we prove the inequality (1). If node $v$ is connected to a neighbor $u$ only by edges labeled with dimensions in $D$ then in both $A$ 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; lastly, 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. Thus, we have that $A \leq B$.

Now, we prove the inequality (2). If 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; lastly, 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. Thus, we have that $B \leq C$. □

### *3.3. Measuring the Hubbiness*

We now formally define the concept of multidimensional hub and a possible characterization for it.

**Definition 8 (Multidimensional Hub).** *Let $v$ be a node and $D$ a set of dimensions in a multidimensional network. Given a threshold $\delta$ the node $v$ is a multidimensional hub in the set $D$ iff $Neighbors(v, D) \geq \delta$.*

In general, the threshold $\delta$ depends on the specific network, although there are empirical rules in the literature to determine it (one example is the classical 80-20 rule [22]). This is why hereafter we omit this threshold, saying only that a hub is a node with a high number of neighbors.

At this point, one question arises: can we give a formal characterization of multidimensional hubs? The set of measures to assess the relevance of a dimension for a given node allows to characterize some kind of hubs. In particular, by combining the two following notions of multidimensional hub and relevance of a dimension for a node we are able to identify, within a set of multidimensional hubs, those for which a specific dimension $d$ is relevant (Definition 9) or irrelevant (Definition 10).

**Definition 9 (D-supported Hub).** *Let $v$ and $D$ be a node and a set of dimensions in a multidimensional network, respectively. The node $v$ is D-supported if $v$ is a multidimensional hub with respect to a set of dimensions $D'$, such that $D \subseteq D'$, and $R(v, D) \geq \epsilon$, with $R \in \{DR, DR_{XOR}, DR_W\}$*

**Definition 10 (D-unsupported Hub).** *Let $v$ and $D$ be a node and a set of dimensions in a multidimensional network, respectively. The node $v$ is D-unsupported if $v$ is a multidimensional hub with respect to a set of dimensions $D'$, such that $D \subseteq D'$, and $R(v, D) \leq \epsilon$, with $R \in \{DR,\ DR_{XOR},\ DR_W\}$*

As one can see, as the difference between the two resides only in the direction of the inequality, each of the two concepts acts as *nemesis* for the other one, hence we use this term hereafter to refer to hubs that play the opposite role of other ones.

There are two *caveats* in the above definitions. First, the definitions are generic for any set of dimensions $D$, where $D$ might even contain only a single dimension. When analyzing real networks, a specific target of analysis might be to find the set of d-supported hubs for one single specific dimension $d$. For an analogous reason, $D'$ might be any set of dimensions included in $L$, in which $v$ is a hub. In our examples in Section 5 we take $D' = L$, as we want to take into account all the available dimensions.

Second, the choice among the various DRs allows to find D-supported (D-unsupported) hubs with very different multidimensional characteristics. The choice is ad-hoc, and only depends on the analysis that one might want to perform, hence there is no better choice among the others. For example, by choosing the $\mathrm{DR}_{XOR}$, and looking for the d-unsupported hubs for a specific dimension $d$, we are looking for hubs that would be hubs even without the connections provided by dimension $d$.

Note that the above characterization in a network whose set of dimensions $L$ would contain only a single dimension $d$, would not make any sense: all the involved sets ($L$, $D$, and $D'$) would contain only $d$, thus (1) all the values for the DRs would be 1, making the distinction between D-supported and D-unsupported vain, and (2) there would be no distinction among the three DRs, making thus the characterization leading to only one possible type of hubs, which is, obviously, the traditional concept of monodimensional hub.

Given all the above, building a multidimensional analysis aimed at extracting and characterizing a multidimensional hub is relatively easy: the analyst defines the desired analysis, translates it in terms of a filter on the values of Dimension Relevance and then selects, among the nodes with high number of neighbors, the ones satisfying the filter, leading to D-supported or D-unsupported hubs, according to the most appropriate choice of DR and parameters.

**Example 1 (Airline Network).** *Without looking at the complete structure of the multidimensional network of airlines (each airline company taken as a dimension), we selected two European multidimensional hubs ($\geq$ 100 connected cities): Dublin and Madrid. We found that the Ryanair airline has a DR of 0.54 for Dublin and 0.27 for Madrid, while it has a $DR_{XOR}$ of 0.31 for the former, and 0.09 for the latter. This means that, while the Ryanair's importance seems to be double for Dublin w.r.t Madrid in terms of connected cities, its importance as sole connection is more than triple. Dublin is then a Ryanair-supported hub, according to both DR and $DR_{XOR}$.*

**Algorithm 1** $MHA - MultidimensionalHubAnalysis$

**Require:** $V, E, D \subseteq L$
**Ensure:** statistics for all nodes in V, w.r.t D

```
1: for all e ∈ E do
2:    increaseNeighbors(srcNode(e), trgtNode(e))
3:    for all d ∈ dimensions(e) do
4:       increaseNeighbors(srcNode(e), trgtNode(e), d)
5:       updateDR_W(srcNode(e), trgtNode(e), d)
6:       if number of dimensions of e is 1 then
7:          increaseNeighbors_XOR(srcNode(e),trgtNode(e),d)
8:       end if
9:       updateDR_XOR(srcNode(e), trgtNode(e), d)
10:   end for
11: end for
```

### *3.4. Implementation and Complexity*

Algorithm 1 is the pseudo-code for computing our measures. Assuming the list of edges to be sorted (this can be done once for all), each measure can be computed by a single scan. In line 2 we update, edge by edge, the Neighbors. In lines 3-10 we scan the dimensions in which each edge appears: in lines 5 and 6 we update the Neighbors with respect to each dimension $d$, and the $DimensionRelevance_W$. Then we check whether we have to update $Neighbors_{XOR}$. Then in line 9 we update the $DimensionRelevance_{XOR}$.

Under the assumption of having a sorted edge list, for each node we keep in main memory the following information: an integer for the number of neighbors, an integer for each dimension representing the degree of that node in that dimension, two floats for each dimension representing the Neighbor XOR and a temporary value for computing the Weighted Dimension Relevance. As soon as the source/destination pair changes, we can release the temporary variables.Thus the space complexity is $O(|N|\times|L|)$, that can be considered as $O(|N|)$, since usually $|L| \ll |N|$. Given that each edge is scanned exactly once, the time complexity for computing the complete set of measures is $O(|E|)$. Sorting the edges can be done once for all when preparing the network, thus we ignore the additional complexity of $O(|E| \times \log(|E|))$.

## 4. Evaluation of Multidimensional Measures

We now want to answer the following:

**Q1.** Are the presented multidimensional measures able to make important latent knowledge emerge from the data?

**Q2.** Would it be possible to extract (part of) this knowledge with non-multidimensional techniques with the same degree of complexity?

**Q3.** What kind of knowledge would the measures make emerge on null models?

We address Q1 in Section 4.1, Q2 in Section 4.2, and Q3 in Section 4.3. In order to do so, we analyzed hubs in the three different real-world networks presented in Section 2: Flickr, QueryLog and DBLP. All the experiments were conducted on a PC with a Core2 Duo processor at 2GHz with 3GB of RAM, running Linux Ubuntu9.10. The binaries of the Java classes and the data used in the paper are available online [4] . In line with linear time complexity, the running times were less than one hour for Flickr, and less than two minutes for QueryLog and DBLP. The memory occupation was less than 1 GB for Flickr and less than 500MB for QueryLog and DBLP.

### *4.1. Multidimensional Measures on Real Networks*

Here, we want to study the power of our multidimensional tools in letting latent knowledge emerge from the data. Figures 4(a)-(c) show, for the three datasets, the cumulative neighbor distributions in log-log scale. Consider the curve corresponding to the global network, i.e. the distribution of neighbors computed over all the dimensions. The DBLP network shows a behavior similar to the "the rich gets richer", with very different cut-offs, while the other networks behave differently. The figures show that the behavior of this measure resembles the one of the degree in the monodimensional setting, even without being completely similar. To support this, in Figure 4(a)-(c) we report also the cumulative neighbor distribution per dimension (which, in turn, is the degree per dimension) of the three networks, and we compare them with the global neighbors distribution. For DBLP, we chose only six representative dimensions out of the original 65.

In Figure 4(d)-(l) we report the distributions of Dimension Relevance in the three dataset. The strong differences among the three networks highlight the presence, in the real world, of networks with different multidimensional structure.

We then believe that the three DRs are able to make the interplay among the dimensions emerge from the data, extracting the knowledge at the center of investigation in Q1, that we now consider successfully answered.

### *4.2. Finding multidimensional hubs with monodimensional techniques*

In order to answer the question "can we extract multidimensional hubs with monodimensional techniques?", the first question to answer is "are multidimensional hub necessarily monodimensional and vice versa?"

Table 2 answers this for our three networks. For each dataset we extracted the top 20% monodimensional hubs (nodes with a high degree in one dimension) and the 20% multidimensional hubs (only taking into account the total number of neighbors considering all the dimensions). The columns of the table report the probability of being a multidimensional hub given that a node is a monodimensional hub and vice versa. We can see from DBLP and Flickr dataset that being a monodimensional hub does not entail being a multidimensional hub and vice versa.

| Network | Multi → Mono | Mono → Multi |
|---|---|---|
| QueryLog | 75.69% | 99.85% |
| Flickr | 70.87% | 46.43% |
| DBLP | 31.08% | 70.87% |

Table 2: Relationship between mono and multidimensional hubbiness of a node

However, one can argue that finding 46% of multidimensional hubs by extracting monodimensional hubs could be sufficient. To prove that this is not true, we show that two multidimensional hubs may look very different when their multidimensional connectivity is examined, or, in other words: the fact that two hubs are multidimensional does not entail that these two nodes have the same importance and show the same behavior. This is based on the intuition that, in the multidimensional setting, two different multidimensional hubs may exhibit a different interplay among the dimensions in which they appear. In order to show this, we report in Figure 4(m-o) the cumulative standard deviation of the three measures for each hub on the different dimensions. The high values of the standard deviation obtained highlight a high diversity of relevance for each of the dimensions in which a node is connected. All the networks show high values of these metrics for a large fraction of nodes. As a result, two multidimensional hubs may look very differently when their multidimensional connectivity is examined.

Consider Figure 5. Here we report the size of the overlap among two sets of hubs: the ones extracted with our filter defined in Section 4.1 and the ones having only a high monodimensional degree. Note that the set of hubs extracted in our analysis here is a subset of the total set of multidimensional hubs. Therefore the set of nodes used for Figure 5 is higly differs from the one used for Table 2. The overlap between the two sets is computed after increasing the number of hubs extracted from the network. We started extracting the 0.25% of high degree nodes and we ended extracting the 2.5% top hubs. The plot highlights two different things. The first is related to Flickr and QueryLog datasets. In these datasets it is fairly impossible to extract the desired set of hubs, answering to our precise analytical questions expressed in Section 4.1, without any multidimensional information. In order to extract less than 1% of the nodes with the desired multidimensional properties, the analyst must extract the 2.5% of the network's hubs. This means, for example, that in order to obtain 7 hubs in the QueryLog dataset the analyst has to extract 5000 hubs and for 200 Flickr hubs this number raises up to 30000. Furthermore there is no way to distinguish the desired hubs from the other ones. The DBLP dataset behaves differently. In DBLP we can obtain almost all (99%) the interesting hubs defined ac-

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

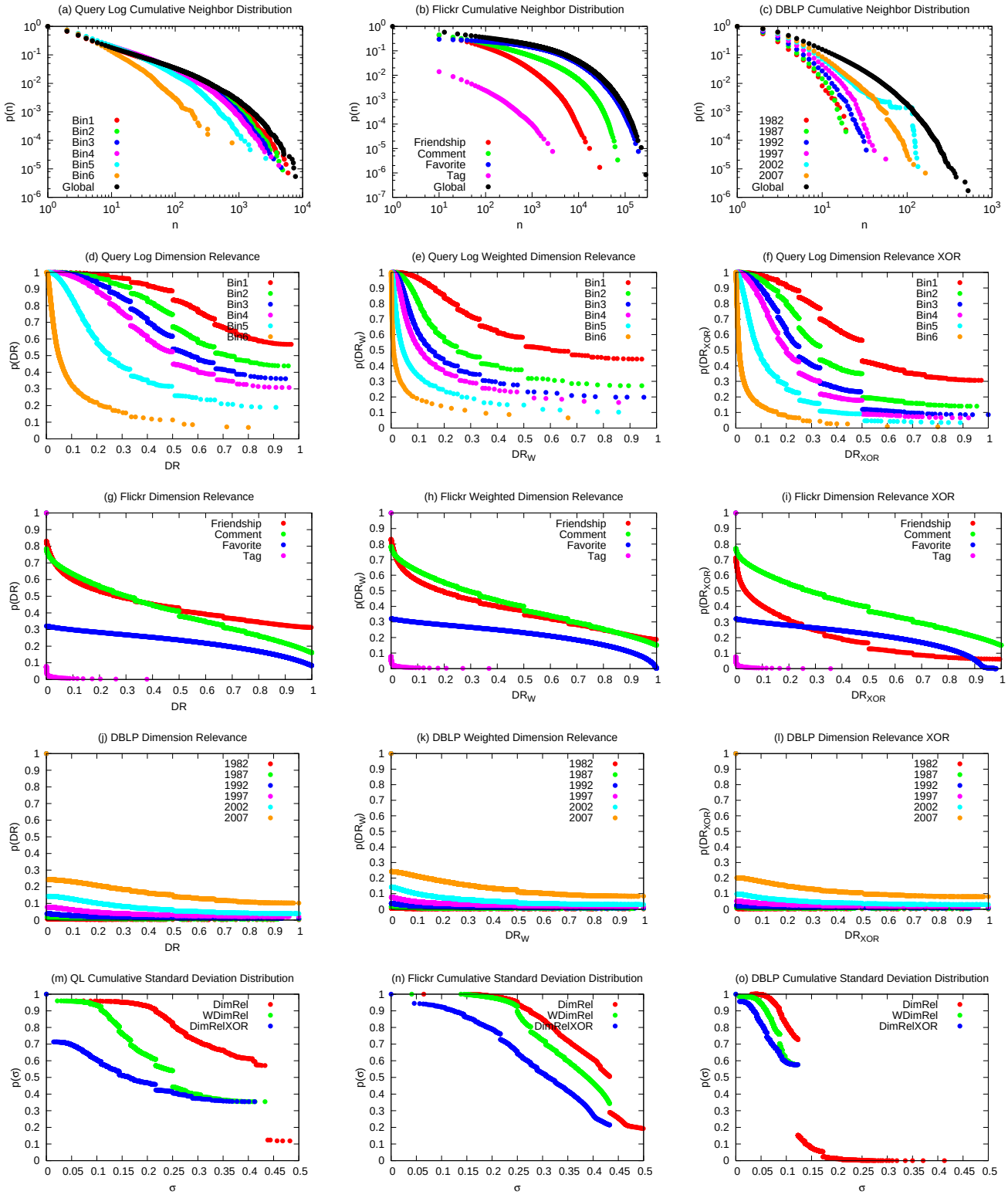



Figure 4: The metrics computed on the three networks (color image).

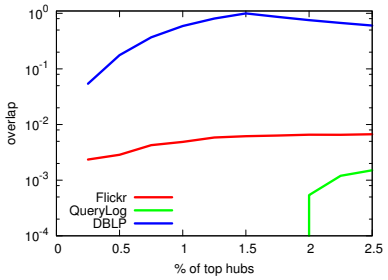



Figure 5: The overlap ratio between monodimensional and multidimensional hubs

cording to our analytical questions by extracting the 1.5% of the hubs of the network (9000 nodes). However, this ratio decreases as we enlarge the set of hubs extracted. This happens because 8774 is the exact number of nodes in DBLP having the desired characteristics. Thus they are not hubs: we are dealing with all the nodes, regardless their connectivity. For this analysis it is a coincidence that all these nodes are also monodimensional hubs, but, as one can expect, this is not always true.

In conclusion, we have provided a motivated answer for question **Q2**, that makes it clear the need for these multidimensional techniques.

### *4.3. Evaluating the measures on null models*

One question left open is: what kind of knowledge would the DRs extract in null models such as a random multidimensional network? Therefore, would their distributions on random networks look similar to the original ones? This point is crucial and would show how the measures are effectively telling something about real, non random, phenomena. Purpose of this section is to study the three DRs under this perspective by means of evaluation of them on different synthetic networks used as null models.

We built four different multidimensional network generators, each with different characteristics, starting from a simple random generator, towards a generator that tries to preserve a global property of the original network that we might see as correlated with our measures, namely the Jaccard correlation index computed among the sets of edges corresponding to the dimensions. For each model, we present its characteristics and the evaluation of the DRs on the QueryLog and the DBLP networks.

Note that, while our measures can be computed with low time complexity, executing the generators might require quadratic (or even more) space and/or time and does not scale well. For this reason, while we can efficiently handle large networks with our measures, we only computed the null models on the smallest ones.

#### *4.3.1. Random*

We created a generator of random multidimensional networks, which takes in input the number of dimensions to generate, and the number of nodes and edges to put into each dimension. We fed the generator with these statistics computed on the real QueryLog and DBLP networks.

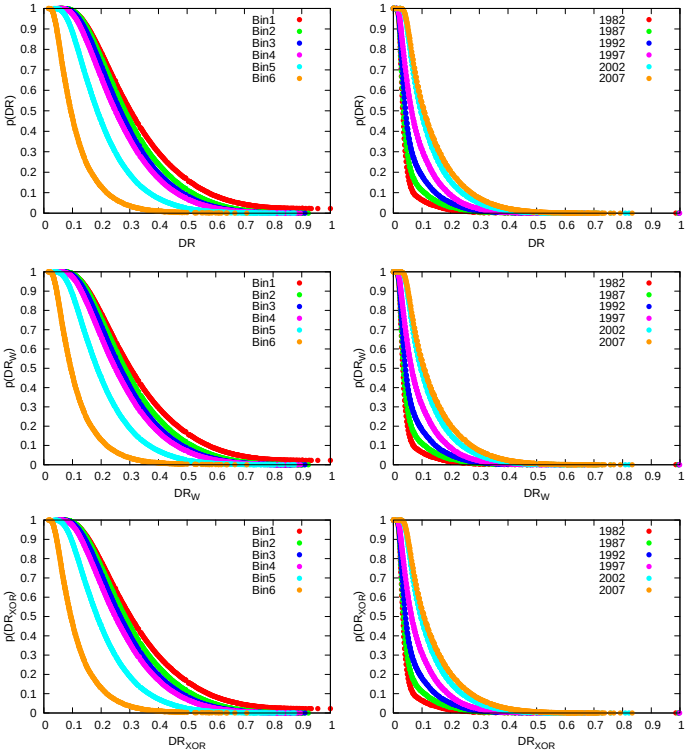



Figure 6: Random: QueryLog (left column) and DBLP (right)

Figure 6 shows the cumulative distribution of the $DR$ (top row), $DR_W$ (central row), and $DR_{XOR}$ (bottom row), computed on the QueryLog-like (left column) and DBLP-like (right column) networks. As we expected, the distributions of the DRs looks much different with respect to the original ones, and the relationships among the dimensions residing within the original networks look destroyed when compared to the original distributions (see Figure 4(d-f) for QueryLog and Figure 4(j-l) for DBLP). Note that the distributions per dimension do not overlap, as we might expect for a random graph, given that we are preserving the number of nodes and edges per dimension, and this causes the DRs computed for each dimension to take different values.

The distributions prove that the knowledge extracted by the DRs on random networks is much different with respect to the one deriving from real data, thus making the knowledge extractable with this analysis on real data non random, supporting then the meaningfulness of the measures.

We then wanted to see in the next generators what we can add to the null model in order to make the distribution look closer.

#### *4.3.2. Preferential attachment*

For the second generator, we took in input the same parameter as the previous one, but we built every dimension by evolving it following the preferential attachment model [4], i.e., after a bootstrap consisting of a clique of three nodes, we iteratively added a node attaching it to a random node with a probability directly proportional to

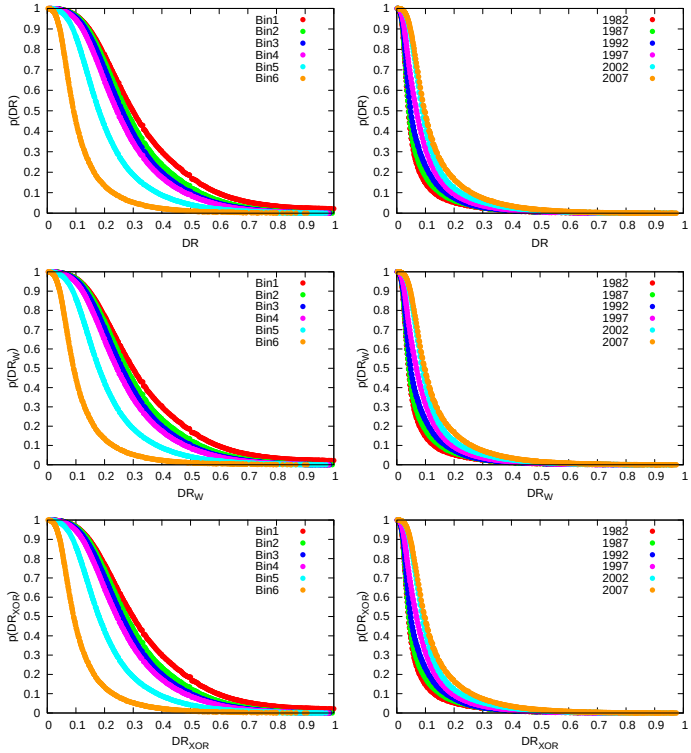



Figure 7: Preferential attachment: QueryLog (left column) and DBLP (right column)

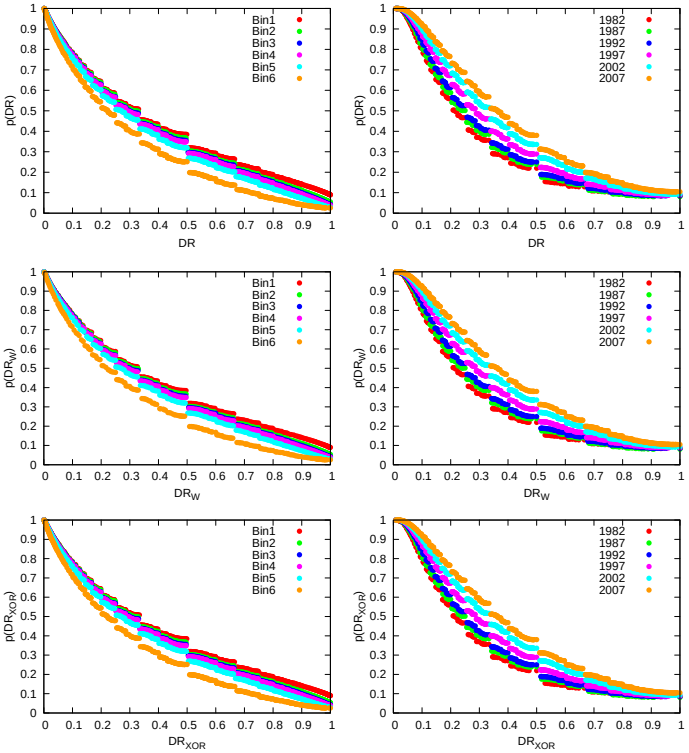



Figure 8: Shuffle: QueryLog (left column) and DBLP (right column)

its degree. Figure 7 reports the distributions of the DRs computed on the two networks. As we can see, we are not adding any significant information to the model compared to the random graph.

### *4.3.3. Shuffle*

The previous two generators are however producing random combinations of links, which is, obviously, destroying most of the original information. In this generator, instead, we keep, dimension by dimension, all the characteristics of the original graph, except the relations among the dimensions. More clearly, we split the graph by dimensions, and we re-merge them in a random way, shuffling then all the node id correspondences among different dimensions. In this way, except destroying the interplay among dimensions, we are keeping most of the characteristics of the original networks.

As one can see in Figure 8 we had results similar to the previous ones. At this point we might think that there is a strong relationship between the global correlation among dimensions, and the values of the DRs, that are, however, local measures.

### *4.3.4. Jaccard*

In order to validate the above hypothesis, we built a generator that preserves only the Jaccard correlation coefficient among the dimensions, computed on the sets of edges of each pair of dimensions of the real network in input. To be more clear, for each pair of dimensions $x$ and $y$ in the original network, we computed $\frac{|E_x \cap E_y|}{|E_x \cup E_y|}$, where $E_x$ and $E_y$ are the sets of edges belonging to dimensions $x$ and $y$ respectively, and generated a network preserving all these values. This was achieved by storing the set of multiedges connecting two nodes, and by using them to build the synthetic graph. The aim of this generator is to preserve the global interplay residing among the dimensions.

As Figure 9 shows, for QueryLog we are now a little closer to the original distribution of the pure DR, while this does not hold for the other two measures, nor for DBLP. This is not surprising, as, by its definition, the capability of the pure DR to capture the interplay among the dimensions is weaker with respect to the other two. In particular, the exclusivity of the $\mathrm{DR}_{XOR}$ is a stronger concept, which is harder to preserve by this generator.

A different consideration must be done to explain the results in DBLP. Looking at figures 4(j-l), we see how the distributions of the three measures are changed in these synthetic networks, in contrast to what happens to QueryLog. However, even though for sake of simplicity we plot only six of them, DBLP has a total of 65 dimensions, thus making it more difficult to preserve the interplay among all of them, even with a little perturbation of the real network. This effect is weaker in QueryLog, that has a total of six dimensions.

### *4.3.5. Conclusions on null models*

Based on our experience matured on null models, we can claim that the DRs are indeed capturing the intrinsic, real, relationships among the dimensions. The phenom-

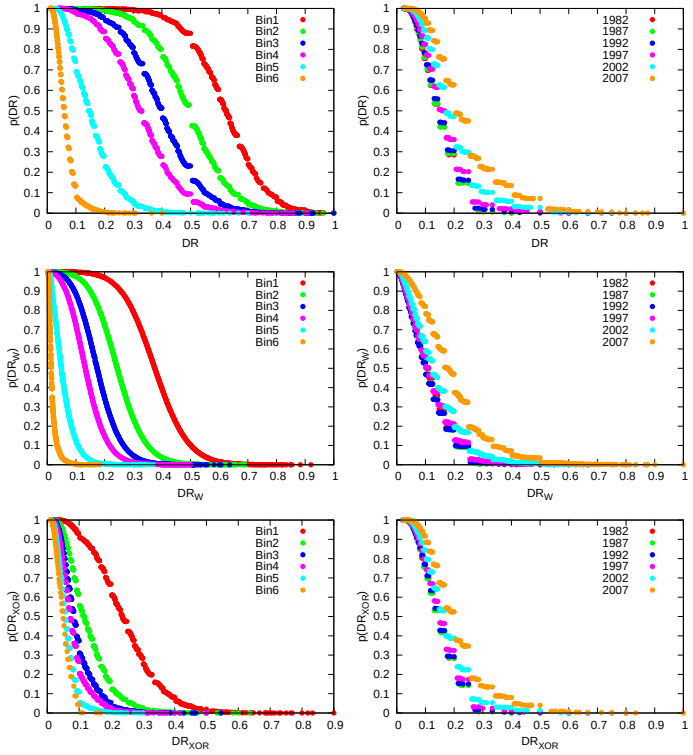



Figure 9: Jaccard: QueryLog (left column) and DBLP (column)

ena caught by these measures are not random, and cannot be easily reproducible with a generator of low complexity. Further investigation might lead to a generator capable of a more sophisticate synthesizing of the values taken by the measures, but at the eventual cost of additional complexity. In addition, it appears clear that the DRs work at the local level and, while it is relatively easy to build a generator based on global properties, the definition of a model preserving the global distribution of a local measure is non-trivial.

This adds to the previous section, where we have showed that it is not possible to measure the same relationships by means of only monodimensional techniques. Based on these two sections then, we believe there is strong motivation behind the use of such measures, whose semantic meaning is unique and justifies their need.

In conclusion, we believe to have successfully answered Q3 by means of the above analysis on null models.

## 5. Hub Characterization in Real Networks

In this section we show how, by exploiting the characteristics and the semantic of the real networks described in Section 2 and of their dimensions, we are able to assign a name to some of the possible characterizations of the hubs. We then find hubs that are interesting w.r.t three simple analytical examples. In our networks, we found convenient to use the dimension relevances as a powerful filter for characterizing a narrow set of hubs, due to the distributions of these measures. In particular, in QueryLog, only 100 hubs have a Weighted Dimension Relevance lower than 0.25 or higher than 0.5. The vast majority of hubs lays on a very narrow interval of values, thus becoming clearly irrelevant for the analysis, more focused on the outliers. This holds also for the other two networks.

The following three examples are meant to be only a sample of possible real-life applications in which our techniques may be helpful. In the future, we plan to expand the direction of finding interesting real-life problems in multidimensional network analysis, in which our techniques might be used as a support for a more complete understanding of real phenomena. Just to give an example of this, we will very briefly present also the *nemesis* (see Section 3) of our extracted hubs, i.e. hubs with very similar number of neighbors, but extracted with a specular filter on the Dimension Relevance. This will help to better characterize the extracted hubs and will give a further idea of the degree of freedom of the analyst in using these analytical tools.

In the following, we consider hubs the nodes with a high number of neighbors taking into account the complete set of available dimensions, i.e., in definitions 9 and 10, we put $D' = L$.

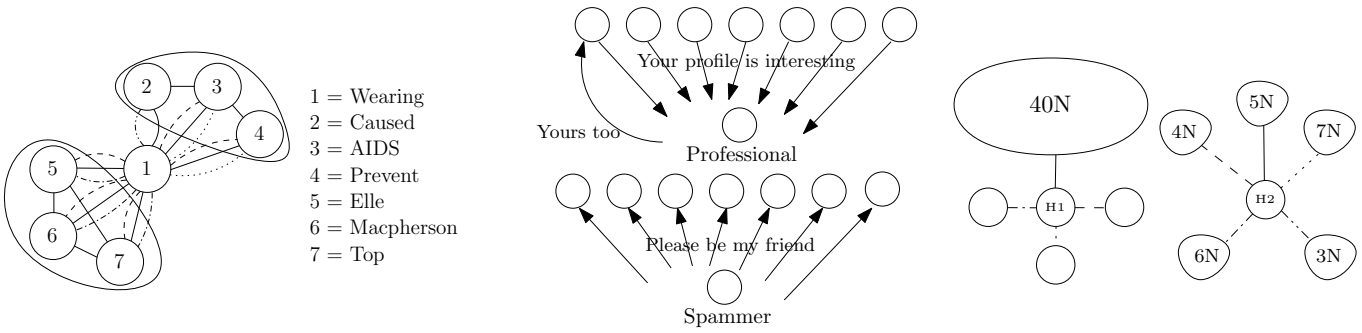
### 5.1. Detection of Ambiguous Query Terms

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

Then, we are saying that being an ambiguous query term translates into being a D-irrelevant hub, where $D = \{\text{“Bin1”}\}$ and the proper dimension relevance measure is the $DR_W$. Note this choice: minimizing the $DR_{XOR}$ of dimension “Bin1” would have selected terms that generally do not produce good results at all, while the pure $DR$ would not have specified the interplay with the other dimensions.

Given the hubs extracted with the above characterization, we wanted to go further, trying to explain why the terms led also to good results in a few cases. We then considered the small communities of words surrounding the hubs extracted, where we looked for the reasons for a very good or very bad query result. We selected the neighbors with the highest Dimension Relevance for dimension 1, to see why, with a generally bad query term, sometimes we find good results.

A possible example found to satisfy these criteria is the word “Wearing” (a simplified view of its neighborhood is depicted in Figure 10a). This term shows here poor semantics, which needs a disambiguation. Moreover, the clusters surrounding this word are very clear: words in either cluster are not really expected to be in the other




(a) (b) (c)

Figure 10: Some of the multidimensional hubs extracted

one. The first group of queries was apparently generated by users looking for information about AIDS and how to prevent it. In the second cluster we see people interested in Elle MacPherson's dressing habits.

The nemesis of this hub, i.e. words which always lead to good results with a very high number of other words (D-supported hub, where $D = \{$"Bin1"$\}$ using $DR_W$), are the words "Wikipedia" and "Amazon": a possible explanation for this is that a user looking either for many different words in an encyclopedia or for products in a store is likely to find the first results to be the best matches.

### *5.2. Outlier Detection*

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

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

Hence, in this analysis, we can characterize as *outliers* the D-relevant hubs, where $D$={"Friendship"} and the dimension relevance is the $DR_{XOR}$.

We wanted to go further, by identifying two subcategories of our outliers: *professionals* and *spammers*, for which Figure 10b gives a possible representation. The first can be identified due to their high number of ingoing edges and the low number of outgoing ones (to do this, we extracted a posteriori the direction of ever edge, distinguish then between ingoing and outgoing ones). This behavior is classic in social networks: if a person has an interesting profile, many people will ask for friendship. We found two instances of this kind of profile[5,6]. On the other hand, the owner of an interesting profile could not be interested in having so many friends. The opposite observation can be made for spammers: they can be detected by a high number of outgoing edges but no one is interested in returning the friendship link to a spammer (we found two examples of these hubs[7,8]).

As nemesis (D-unsupported hubs, where $D$={ "Friendship"} and using $DR_{XOR}$), we found three profiles[9,10,11]. All these profile presented, at the time of the download of the network, a very high number of neighbors and no one exclusively through the Friendship dimension: at the moment of writing this paper, all the profiles are closed. Therefore, the nemesis of both *spammers* and *professionals* are the *quitters* (and this is really interesting in the perspective of the service providers).

### *5.3. Analyzing Temporal Behaviors*

In this section, we go beyond the theory presented so far. Consider definitions 9 and 10. It is clear that real networks might express rich semantic, and that even powerful tools and characterizations as defined so far could not cover the complete set of analyses that one might want to perform. In this perspective, we want to show how, by substituting the usage of the DRs in the two mentioned definitions with any of the possible aggregates computed on their values, it is possible to expand the class of phenomena that can be studied with our tools.

In this context, an interesting application of our approach is to analyze the temporal behavior of multidimensional hubs on evolving networks. In this section we show the results obtained on DBLP, whose dimensions are the years of publications. The specific object of our analysis is to find authors of scientific papers who tend to change the authors with whom they collaborate possibly every year. Note that we are not focusing on just new collaborations, but we want also to see the old ones to disappear. In order to do so, we found hubs $v$ maximizing the number of dimensions $d$ for which $\mathrm{DR}_{XOR}(v, d) > 0$ (maximizing this value means maximizing the number of years in which the

[5] http://www.flickr.com/photos/38687875@N00
[6] http://www.flickr.com/photos/20532904@N00
[7] http://www.flickr.com/photos/10539246@N05
[8] http://www.flickr.com/photos/23941584@N08
[9] http://www.flickr.com/photos/21700048@N04
[10] http://www.flickr.com/photos/22045276@N00
[11] http://www.flickr.com/photos/53654438@N00

author had collaborations that took place only in a specific year and not in others).

In this scenario then, we call then *dynamic researchers* the D-relevant hubs $v$, where $D = L$ (where $L$ contains all the years) and, instead of the any of the simple DRs, we maximize $|\{d : DR_{XOR}(v, d) > 0\}|$.

Figure 10c reports two representations of hubs extracted in this way: the hubs behaving as H1 and the ones behaving as H2. To be more precise, a deeper classification among them might be performed by looking also at the standard deviation of the $DR_{XOR}$ computed in all the dimensions. The example H2 in the right of that Figure, in fact, represents a hub minimizing the standard deviation. H1 hubs are collaborators in high effort publications such as books (such as Maxine D. Brown or Steffen Schulze-Kremer); while H2 hubs are authors who tend work with many different people, rarely keeping these collaborations alive, such as Ming Xu or Jakob Nielsen.

Finally, if we minimize the $\mathrm{DR}_{XOR}(v, d)$ we find the nemesis of these hubs. The list of these hubs includes many relevant names in Computer Science: Allan Borodin, Richard M. Karp, Robert Endre Tarjan, Godfried T. Toussaint, and Jeffrey D. Ullman fall in this category.

## 6. Related work

In this section we briefly review some research related to our analysis from two different points of view: the analysis of hubs, and possible models and measures for multidimensional networks.

Scale-free networks, i.e. networks with the degree distribution following a power law, have been studied for many years. The first study introducing the term "scale-free" was [4], where the authors discovered that the structure of the Web shows the presence of a few highly connected nodes, the hubs, and many nodes with a low degree. Other papers studied the same concept and tried to capture the "importance" of a node in a network: [14] is a well known example. Since then, many papers have considered scale-free networks in several different areas of research. In [8], the authors analyzed the spread of viruses in real networks, showing that the best nodes to immunize in order to prevent the spread are not necessarily hubs. In social networks, many studies have analyzed the power of highly connected and influential nodes from different points of view: [11, 24, 7] are just a few, describing how having highly connected nodes affects the social behavior of the networks. An interesting study on in citation and collaboration networks is presented in [29], where the authors use heterogeneous networks, which can be considered very similar to our multidimensional setting. In communication networks, the authors of [1] showed how to make use of hubs in peer-to-peer networks for fast and efficient searches. In relation to hubs in networks it is impossible not to mention previous approaches like PageRank [19] or HITS [14]. However, there are substantial differences with our setting that an experimental comparison would not make much sense of. First, we are analyzing multidimensional networks, while both of these studies were proposed in the monodimensional setting. Secondly, although our approach could be extended in order to deal with directed edges, in this paper we handle undirected networks. Finally, our measures are local to the hub, and do not consider any kind of hubbiness inherited from the neighbors.

As we can see, all the previous methods disregard the possibility of enhancing their analysis with the power of a multidimensional investigation, which can be extended in order to consider this more complex scenario.

Nevertheless, in the last couple of years, a few studies have been proposed in order to capture the natural multiplicity of relationships. Some research focuses on specific multidimensional social networks, such as communication networks among people. In one of these papers [26], the aim was not the analysis of hubs but the multidimensional formulation of machine learning tasks on social network. Given a network and a set of latent social dimensions the authors were able to determine how new entities will behave in these dimensions. Another interesting paper treating multidimensional networks is [9], which introduces the *graph OLAP*, a multidimensional view of graph data. The paper defines *informational* and *topological* dimensions over a graph, which correspond simply to different observations of the same graph and its different hierarchical views.

Two more papers deal with the analysis of multidimensional network [25, 16]. In both cases, the authors analyze networks with "positive" and "negative" links among online communities. The authors in [25] analyze the degree distributions of the various dimensions, which are scale-free structures, highlighting the need for analytical tools for the multidimensional study of hubs. In [16], the authors focus on link prediction in multidimensional networks.

Other studies focus on the analysis of multi-relational networks aimed at capturing the variety of relationships between different entities [10, 23]. Most of these works, however, only consider the possibility for nodes to be connected via different kinds of relationships, while the presence of multiple connections *at the same time* is mostly disregarded.

In summary, a definition of new analytical measures is lacking, and the interplay among different dimensions has not been investigated in any way. In our previous work [6] we tried to overcome to this, by defining a model and a full framework for multidimensional network analysis.

## 7. Conclusions

In this paper, we have addressed the problem of identifying and characterizing multidimensional hubs in real world networks by defining suitable analytical tools.

We applied our scalable methodology to large real networks and showed that such hubs do exist and they can

be found and studied by using our measures of interplay of the different dimensions. Moreover, our measures allow to discover and quantify the importance of every single dimension above the others.

While pursuing the research illustrated in this paper, we learned that analyzing multidimensional networks is an interesting research direction, which opens a variety of new questions and requires the definition of new analytical tool, such as the dimensional relevance measure introduced here; we are currently investigating a comprehensive repertoire of multidimensional network measures, including distance, centrality, clustering and so on: [6] is a technical report describing the current state of our broader framework for multidimensional network analysis.

In fact, many other questions on multidimensional networks remain unanswered, and call for further research; we mention two such lines briefly here.

First, we did not consider, in our approach, the possible structure or semantics of the specific set of dimensions under analysis: each different dimension is a distinct categorical value, and used as such in the multidimensional measures; however, such dimension values can be meaningfully sorted (as, e.g., in the QueryLog network, where dimensions are associated to quality levels) or may have a temporal or spatial semantics (as, e.g., in the DBLP network, where dimensions are associated to years). How can our measures be extended to fully exploit this additional structure?

Second, it would be interesting to devise a generalized query framework for the discovery and analysis of hubs in multidimensional networks, based on the proposed measures, capable of supporting the analyist in expressing the desired queries (e.g., top-k hubs according to some specified hubbiness and relevance constraints), in finding appropriate parameters and thresholds for the involved measures on the basis of the available network data.

Finally, as mentioned in Section 2.2, we plan to extend our study on other kinds of centrality, such as betweeness, closeness or eigenvector centrality.

## Bibliography


[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, Search in power-law networks, CoRR cs.NI/0103016 (2001).

[2] D. Balcan, B. Gonalves, H. Hu, V. Colizza, J. J. Ramasco and A. Vespignani, Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model, Journal of Computational Science 1 (2010) 132-145.

[3] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47–97.

[4] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509.

[5] M. Berlingerio, F. Bonchi, B. Bringmann, A. Gionis, Mining graph evolution rules, in: W. L. Buntine, M. Grobelnik, D. Mladenic, J. Shawe-Taylor (Eds.), ECML/PKDD (1), volume 5781 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 115–130.

[6] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi, Foundations of multidimensional network analysis. Tech. Rep. 2010-TR-004, `http://puma.isti.cnr.it/dfdownload.php?ident=/cnr.isti/2010-TR-004`, 2010.

[7] K. M. Borgwardt, H.-P. Kriegel, P. Wackersreuther, Pattern mining in frequent dynamic subgraphs, in: ICDM, IEEE Computer Society, 2006, pp. 818–822.

[8] D. Chakrabarti, Y. W. 0008, C. Wang, J. Leskovec, C. Faloutsos, Epidemic thresholds in real networks, ACM Trans. Inf. Syst. Secur. 10 (2008).

[9] C. Chen, X. Yan, F. Zhu, J. Han, P. S. Yu, Graph olap: Towards online analytical processing on graphs, in: ICDM, IEEE Computer Society, 2008, pp. 103–112.

[10] C., Deng and S., Zheng and H., Xiaofei and Y., Xifeng and H., Jiawei, Community Mining from Multi-relational Networks, in: PKDD, Springer Berlin / Heidelberg, 2005, pp. 445–452.

[11] D. W. Franks, J. Noble, P. Kaufmann, S. Stagl, Extremism propagation in social networks with hubs, Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems 16 (2008) 264–274.

[12] H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.

[13] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabasi, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.

[14] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (1999) 604–632.

[15] J. Leskovec, L. A. Adamic, B. A. Huberman, The dynamics of viral marketing, in: J. Feigenbaum, J. C.-I. Chuang, D. M. Pennock (Eds.), ACM Conference on Electronic Commerce, ACM, 2006, pp. 228–237.

[16] J. Leskovec, D. P. Huttenlocher, J. M. Kleinberg, Predicting positive and negative links in online social networks, in: M. Rappa, P. Jones, J. Freire, S. Chakrabarti (Eds.), WWW, ACM, 2010, pp. 641–650.

[17] A. S. Maiya, T. Y. Berger-Wolf, Online sampling of high centrality individuals in social networks, in: M. J. Zaki, J. X. Yu, B. Ravindran, V. Pudi (Eds.), PAKDD (1), volume 6118 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 91–98.

[18] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J. Onnela, Community Structure in Time-Dependent, Multiscale, and Multiplex Networks, Science 328, 876 (2010).

[19] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, 1998.

[20] G. Pass, A. Chowdhury, C. Torgeson, A picture of search, in: X. Jia (Ed.), Infoscale, volume 152 of *ACM International Conference Proceeding Series*, ACM, 2006, p. 1.

[21] S. Redner, How popular is your paper? an empirical study of the citation distribution, The European Physical Journal B - Condensed Matter and Complex Systems 4 (1998) 131–134.

[22] Newman, M. E. J., Power laws, Pareto distributions and Zipf's law, Contemporary Physics 46 (2005) 323–351.

[23] M. A. Rodriguez and J. Shinavier, Exposing multi-relational networks to single-relational network analysis algorithms, Journal of Informetrics 4 (2010)29–41.

[24] X. Shi, B. Tseng, L. Adamic, Looking at the Blogosphere Topology through Different Lenses, in: ICWSM 2007, volume 1001, p. 48109.

[25] M. Szell, R. Lambiotte, S. Thurner, Trade, conflict and sentiments: Multi-relational organization of large-scale social networks, arXiv.org, 1003.5137 (2010).

[26] L. Tang, H. Liu, Scalable learning of collective behavior based on sparse social dimensions, in: D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, J. J. Lin (Eds.), CIKM, ACM, 2009, pp. 1107–1116.

[27] X. Yan, J. Han, gspan: Graph-based substructure pattern mining, in: ICDM, IEEE Computer Society, 2002, pp. 721–724.

[28] H. Zheng, H. Wang, F. Azuaje, eNelator: A simulation system for large-scale vulnerability analysis of species-, disease- and process-specific protein networks, Journal of Computational Science 1 (2010) 197-205.

[29] D. Zhou, S. A. Orshanskiy, H. Zha, C. L. Giles, Co-ranking authors and documents in a heterogeneous network, in: ICDM, IEEE Computer Society, 2007, pp. 739–744.