Multidimensional Analysis of Linguistic Networks

Tanya Araújo and Sven Banisch

Abstract. Network-based approaches play an increasingly important role in the analysis of data even in systems in which a network representation is not immediately apparent. This is particularly true for linguistic networks, which use to be induced from a linguistic data set for which a network perspective is only one out of several options for representation. Here we introduce a multidimensional framework for network construction and analysis with special focus on linguistic networks. Such a framework is used to show that the higher is the abstraction level of network induction, the harder is the interpretation of the topological indicators used in network analysis. Several examples are provided allowing for the comparison of different linguistic networks as well as to networks in other fields of application of network theory. The computation and the intelligibility of some statistical indicators frequently used in linguistic networks are discussed. It suggests that the field of linguistic networks, by applying statistical tools inspired by network studies in other domains, may, in its current state, have only a limited contribution to the development of linguistic theory.

1 Introduction

Network analysis is an integral component in the study of complex systems. This is probably due to the generally accepted fact that complex systems are composed of elementary units and structures of mutual dependencies between those units which

Tanya Araújo
ISEG (Lisboa School of Economics and Management) of the University of Lisbon and Research Unit on Complexity in Economics (UECE), Portugal
e-mail: tanya@iseg.utl.pt

Sven Banisch
Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22,
D-04103 Leipzig, Germany
e-mail: Sven.Banisch@mis.mpg.de

A. Mehler et al. (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, Understanding Complex Systems, DOI: 10.1007/978-3-662-47238-5_5

directly suggests a network representation. However, network-based analyses are nowadays quite common also in the analysis of systems where such a network representation is not always that intuitive. There may be many ways in which the elementary units and the links between them are conceived and the choices may depend strongly on the questions that a network analysis aims to address.

Here we discuss that case at the example of linguistic networks. Linguistic networks are characterized by a high level of abstraction compared to networks in other areas of research. While in power grid networks, the world wide web or even gene regulatory networks the nodes and links in between them are directly related to real processes taking place in the system – to electricity flow, web links or respectively biochemical reactions between DNA segments – this is not generally the case in linguistic networks which are often induced or synthesized from a linguistic data set for which a network perspective is only one out of several options for representation.

In this context, our main points are:

1. Although being different tasks, network design/induction and network analysis are strongly interdependent
2. Linguistic networks are special both in the design and in the analytical setting
3. There is a need for a framework for network construction and analysis with special focus on linguistic networks
4. A higher level of abstraction in network induction has an important bearing on network analysis, particularly on the choice of appropriate topological indicators and on the interpretation of their results
   or
5. The higher is the abstraction level of network induction, the harder is the interpretation of the topological indicators used in network analysis

The paper is organized as follows: Section 2 presents our main arguments on the specificity of linguistic networks and on the consequent need for the identification of different abstraction levels when dealing with network design. These abstraction levels make up the first dimension of a framework for network construction and analysis. Section 3 is targeted at presenting the second dimension of such a framework, i.e., the statistical levels of network analysis. It also includes a detailed presentation of the statistical indicators most often used at each analytical level. In Section 4, we discuss on the intelligibility of these statistical indicators when applied to the analysis of linguistic networks. Section 5 provides examples in different fields of application together with a classification of the examples according with the introduced framework. Section 6 discusses the classification presented in Section 5, while Section 7 draws a conclusion on the paper as a whole.

2 Linguistic Networks Are Special

2.1 Three Types of Networks

Linguistic networks are characterized by a high level of abstraction compared to networks in other areas of research. They are usually induced or synthesized from a linguistic data set – typically a series of letters organized into words, sentences, may be paragraphs and so on – that does not obviously call for a network representation. This synthesis involves several design decisions and entities (i.e., words) linked in one design may not be linked in another. Moreover, there is no clear relation between the connections in such a network and the dynamic processes taking place in the system that created the data set. It is, for instance, in most cases not possible to state which kind of flow processes take place along the links of a linguistic network which makes the applicability of network measures that involve implicit assumptions about network flow problematic (see Borgatti (2005) and below). For the purposes of this paper, we differentiate roughly between three levels of abstraction by considering:

1. Abstraction Level-1: real networks of systems composed of elementary units which are explicitly linked or in between which real processes take place,
2. Abstraction Level-2: proximity networks of systems composed of elementary units and a well-defined measure of distance in between these units (shared features, correlation or similarity),
3. Abstraction Level-3: induced networks that are synthesized out of data bases (probably the outcome of a complex system) and in which the definition of elements and links is not explicit in the data.

Besides being of great importance at the very beginning of the network definition, the question whether a network or respectively a system falls into category one, two or three has, we believe, an important bearing on network analysis, particularly on the choice of appropriate topological indicators and on the interpretation of the results. As Borgatti (2005) has shown for centrality measures, the implicit assumptions about network flow that certain measures make, may challenge the interpretability of the measure. Namely, most of the well-known network structure indicators have been designed in order to describe the different aspects of transport phenomena where connectivity plays the important role and the mass conservation principle features the dynamics of network flows. Therefore, the situation is even worse if the dynamical processes between nodes are unspecified, as in linguistic networks.

The aim of describing networks at different levels of abstraction is not to present a rigorous classification of different systems or representations of systems. It is rather to show that the topological interpretation of linguistic networks is different from more concrete networks, because a rather abstract perspective has to be taken in deriving a network representation which leads to a certain degree of arbitrariness and a decrease in explanatory power.

For the first type of network (henceforward called type-1), it is relatively obvious how elements and connections in between them must be defined in order to obtain an operative description of the system. For instance, the system of air traffic is constituted of airports and flight connections in between them (Li and Cai 2004; Colizza et al. 2007; Opsahl et al. 2010; Konect 2014). A network which contains just that binary information (flight connection or not) can already be quite informative of the entire air traffic system as questions of efficiency, vulnerability etc. can be addressed at this level. Considering the amount of goods traveling from airport to airport or the number of people will quickly lead to a very detailed description of the whole system.

As another economic example, we mention networks of inter-bank dependencies (Huizinga and Nicodème 2004; Boss et al. 2004; McGuire and Tarashev 2006; Soramäki et al. 2007; Minoiu and Reyes 2011; Spelta and Araújo 2012). Payment systems allow banks to move money and securities between banks and other large financial institutions. Daily, a huge amount of capital flows arises from and depends upon the coordination of payments among banks. Such a close coordination engendered by payment systems creates a network of inter-bank dependencies, where banks are nodes and transactions are represented as links of the network. In this context, the main issue uses to be associated to the conditions that ensure network robustness since failures of a bank to make payments can trigger a cascade of liquidity losses.

Another example, already mentioned above, is the WWW (e.g., Albert et al. (1999), Broder et al. (2000), and Meusel et al. (2014). Web pages are connected to one another by hyperlinks and lot about the system can be understood on the basis of such structural information. However, the WWW is also an area in which the second type of network (type-2) can provide a complementary view on the system as a whole. Instead of hyperlinks as a direct reference from page to page, one could look at the similarity between the pages in terms of the information they provide. One way to measure the similarity between two pages is to compare frequencies with which words are used in the two pages, that is, vectorial semantics (Salton et al. 1975; Salton 1989; Landauer and Dutnais 1997) (for a related approach see also the chapter of Masucci et al. in this volume), or to compare the HTML structure of the documents (Mehler et al. 2007). Of course, the resulting similarity measure depends strongly on the features that are used to span the semantic space. Another way to construct a similarity network is to count the number of common references/hyperlinks or the geodesic distance in the type-1 network. That is to say, for networks of type two, there is already a certain amount of freedom in the definition of the network. It might so happen (even though this is not assumed the regular case) that two nodes close to one another in one network representation are distant in another depending essentially on the type of similarity measure used in the construction.

Some networks that go under the label of linguistic networks fall into this second category. For instance, there are a word networks based on the number of shared letters, phonemes or syllables (see, e.g., Soares et al. (2005), Mukherjee et al. (2009), Arbesman et al. (2010), and Yu et al. (2011)), and, at a higher level, there are interlanguage networks based on lexical similarity (see, e.g., Blanchard et al. (2011) and

Serva et al. (2011)). However, the number of ways to induce a linguistic network is huge and the resulting information represented by the network may differ greatly from one network to the other. This is because linguistics networks are often defined on the basis of linguistic data, such as corpora of texts or annotated dialog data so that many linguistic networks are of the third type, with still more freedom in defining a network, compared to the second network type.

Words, for instance, can be linked because they frequently share the same context (this is captured by co-occurrence networks), but the operationalisation of "sharing the same context" is far from unique (see below).

Two words could also be linked because it is likely that I think of the second when I hear the first (free association experiments; see, for instance, Nelson et al. (2004), Borge-Holthoefer and Arenas (2010), and Gravino et al. (2012)) or if I am asked to utter a sequence of words the two will probably appear together (verbal fluency experiments; see, for instance, Storm (1980), Lerner et al. (2009), Goñi et al. (2011), and Iyengar et al. (2012)). Free word association is probably the only example of a linguistic network of type-1. Although the links between any two words in any utterance sequence are not "hardwired", they are directly related to a real process taking place in the system (or the game) of free association. In this context, a sequence of associated words itself is naturally seen as dynamical process since it provides the unfolding of a flow of ideas. There, the network perspective comes naturally and the definition of nodes and links is relatively obvious: words are nodes and any link is defined as a connection between two spontaneously related words.

Still another way to come to word networks is to map the syntactic relations onto the network connections (syntax dependency networks, see Ferrer i Cancho et al. (2004) and Ferrer i Cancho et al. (2007)) or define a word network by considering explicitly the semantic relations (synonymy, hypernomy, etc.) as the connections in between words (Sowa 1991; Miller 1995). And so on. A further degree of freedom in defining a linguistic network concerns the choice of the elementary units. It is not explicit in the data what we should conceive as the nodes in a linguistic network and therefore models at the level of words, but also at the sentence or paragraph level are sometimes constructed.

The way how a network is defined in linguistics depends a lot on the question that is to be addressed by the network study. If one aims at understanding the structure of a language, syntax dependency network or co-occurence networks synthesized from large corpora of texts are natural choices. If the structure of the conceptual space is in the focus, free word association, verbal fluency experiments or explicit semantic relations are a more appropriate source for network induction. The structure of these different kinds of networks may be very different. The fact that so many different perspectives can be taken on the linguistic system may be seen has an indication of its increased complexity compared to other systems. There are several dimensions of analysis.

2.2 Network Induction

Network induction makes reference to the method by which networks are created on the basis of a certain data set or system. In research dealing with network models of type one, the method of "network induction" does not receive particular attention and many researchers may actually wonder about the notion of "inducing" a network. This is because in those systems the network representation is very natural and requires only a small level of abstraction from the real situation. It is relatively clear what should be considered as the elementary units of the system - the nodes - and also the relation between the nodes - the edges - are given either by real "hard-wired" connections or by processes or flows between the elementary units. Design decisions, if any, concern the question whether the situation is best mapped onto a binary graph or a weighted and directed network.

This is already different for networks of the second type. While, in general, it is often clear what to conceive as the nodes in proximity networks, the nature of the connections (despite that they are similar with respect to a chosen set of features) it is not always clear. In functional brain networks, to make an example, different brain regions are linked if they are jointly activated in certain working tasks (see Sporns et al. (2005), Bassett and Gazzaniga (2011), and Menon (2011) and references therein). Even if it is natural to assume that there is exchange of information or signaling between the different regions during the working task, as a matter of fact, all that can be said on the basis of functional brain network is that there is a correlation of activity patterns across the different brain regions. The real processes (that probably exist) are obscured by the network representation and one would have to go to the micro level of neurons and synapses to obtain that information. That is going to the type-1 network.

Take the similarity graph between articles from the "bibliography on linguistic, cognitive and brain networks" compiled by Ramon Ferrer i Cancho (Ferrer i Cancho 2012) as another example for a type-2 network. This graph, one instance of which is shown in Fig. 1, is obtained on the basis of similar words used in the abstract of the articles. Nodes are labeled by the most frequent term in the respective abstract. While the picture illustrates nicely the structure of similarities between the articles - articles on linguistic networks are clustered into different modules depending on whether they deal with semantics, word networks, text systems or networks of languages as a whole and the articles dealing with functional brain networks form another rather independent cluster - it is not really clear whether linked articles indeed refer to one another and there is also no obvious functional relation or process that is represented by a connection. As we will discuss below, the lack of knowledge about the real relations between the elements has important consequences for the choices and interpretations of certain topological indicators.

The example of a similarity network shown in Fig. 1 makes also clear that various design decisions have to be taken in the creation of such a network. First, in this example, we have been interested in the thematic similarities of the publications as opposed to, for instance, stylistic similarities between different authors. Therefore, we disregarded functional words in the computation of the correlations and

considered only words that do not appear in all the abstracts. If instead we were interested in the identification of authors and similarities in their writing style considering functional words only would probably give the better result (see, e.g., Peng and Hengartner (2002)). Second, once the correlation matrix is computed using the reduced ("cleaned") feature set, a network as shown in Fig. 1 is obtained only after thresholding the correlation matrix such that only strong correlations are preserved. This is also a rather decisive design decision to optimize the intelligibility of the system because, in effect, there is at least a small positive correlation between all pairs of abstracts. However, the complete graph resulting from that would not be very informative about the modularity structure in the network of articles. In fact it happens that when networks are induced from distance measures, the issue of deriving a sparse network from the complete one becomes the most important design decision. The less arbitrary choices (or the most endogenously based ones) usually define the threshold as the distance value used in the last step of the minimum spanning tree construction. In so doing we ensure the connectivity is preserved (the resulting network is necessarily connected).

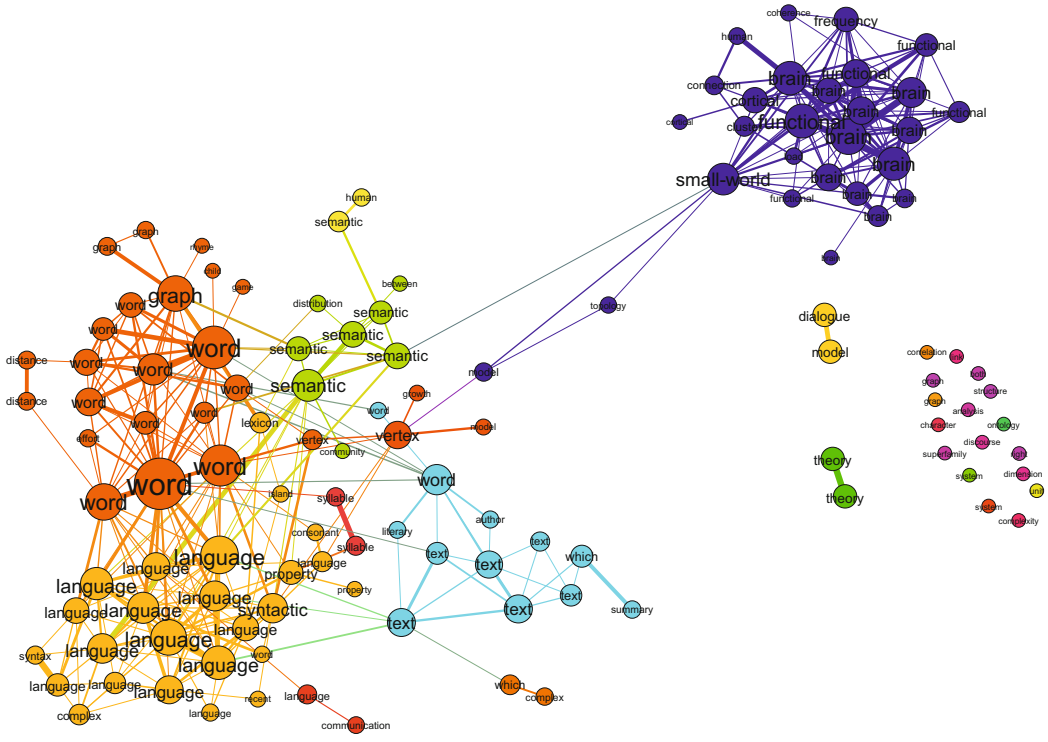

Fig. 1 A similarity network of publications from the "bibliography on linguistic, cognitive and brain networks" compiled by Ramon Ferrer i Cancho (Ferrer i Cancho 2012). The data has been retrieved on the third of December 2012, one week before the conference "Modeling Linguistic Networks" was held. It is constructed by computing the correlation between the abstracts of the articles on the basis of similar words they use. Node labels show the most frequent word in the respective abstract.

The induction process becomes more complex for type-3 networks, and some of the most popular linguistic networks in particular. The number of different options for network induction increases further and several design decisions must be taken from the very beginning of the network creation procedure. Sticking to word networks, one must first be clear about whether phonemic, syntactic, semantic or still other relations should be mapped onto the connections. Usually, this is determined by the research question to be addressed by the network study. Let us assume that the focus is on language structure. Then, we could map syntactic relations between words as they are realized in a set of sentences. This leads to the syntax dependency networks proposed by Ferrer i Cancho et al. (2004). But we could also decide to approach the question by considering shared context or word co-occurrences. Then an important design decision concerns the size of window that is considered as context. Words could be linked whenever they co-occur within at least one sentence, but one could also use a window of fixed size and consider co-occurrence within that window. Another way (considered more realistic in Solé et al. (2010)) is to consider the order in which words occur, most simply, we could say that there is an arrow from word A to word B if B follows A in some sentence (this gives rise to precedence or word-flow networks see, e.g., Grabska-Gradzinska et al. (2012) and below). Again, in all these cases, the question whether links or arrows are weighted or not requires a design decision. In other words, there is a large number of different ways to induce a word network from linguistic data and the data itself does not clearly suggest if one way is better than the other.

The need for a unifying framework for linguistic networks emerges from this diversity of network models. The fact that rather different networks can be obtained on the basis of the same data set seems to be quite unique in the network sciences and triggers, in fact, important epistemological questions concerning validity, comparability and interpretability of linguistic network studies. The comparison of different linguistic networks and a critical discussion of the relation of linguistic networks to networks in other fields are essential for the future development of the field of linguistic networks. First steps into that direction have already been taken (see, for instance, Choudhury et al. (2010), Zamora-López et al. (2011), and Gravino et al. (2012)).

3 Three Levels of Statistical Analysis

Closely related to network analysis and also grounded on a multilevel perspective is the differentiation between statistical levels of analysis (and the corresponding statistical tools). Here we identify three statistical levels inspired by statistical tools originally developed for the characterization of stochastic processes, and their application to the analysis of signals, i.e., signal processing on graphs.

3.1 A Brief Note on Signal Processing on Graphs

Signals evolving in time may be considered as signals in a forward-connected graph, the nodes being different points in time. The analysis of more general networks such as social and economic networks, linguistic networks and biological networks usually generates graphs with much more complex connections.[1]

Processing signals on graphs has been dealt with recently, in particular in the context of discovering efficient data representations for large high-dimensional systems and other dynamical systems (Miller et al. 2010; Shuman et al. 2013). Here we envision that, by analogy with signal processing on graphs, the statistical tools that have been developed for the characterization of stochastic process are mathematical devices that may be applied to the analysis of networks.

3.2 The Statistical Levels

When a phenomenon is measured with a set of statistical tools, what one registers is a sequence of values of some variable X

$$\cdots X_{-2}X_{-1}X_0X_1X_2\cdots$$

which takes values in a space $\mathscr{X}$. The space $\mathscr{X}$ is called *state space* and the space of sequences $\mathscr{X}^{\mathbb{Z}}$ is referred to as *path space*. Statistical properties of the phenomenon may be described at three different levels, (Vilela Mendes et al. 2002):

1. by the expectation values of the observables;
2. by the probability measures on the state space $\mathscr{X}$;
3. by the probability measures on path space $\mathscr{X}^{\mathbb{Z}}$.

To obtain expectation values and probability measures we would require infinite samples and a law of large numbers. For any finite sample we obtain finite versions of the expectation values, of the probability on state space and of the probability on path space which are called the *mean partial sums*, the *empirical measures* (or empirical probability distribution functions - pdf's) and the measures on the *empirical process*.

The statistical levels represent successively finer levels of description of the statistical properties associated to the topological indicators used in network analysis. We will call these three types of description, respectively, level 1, level 2 and 3-statistical indicators.

[1] A first step in analyzing data in such networks is the construction of the appropriate signal transforms. For the forward-connected time graph the Fourier transform is a projection on the eigenvectors of the adjacency matrix. Therefore it is natural to construct transforms for general networks by projection on a basis constructed from the eigenvectors (or generalized eigenvectors) of some matrix relevant to that network.

1. Statistical Level-1 concerns the computation of the quantities related to averages values of one or more topological coefficients defined at the node level, as for instance: the average clustering coefficient, the network degree, the average path length, among others. At this level, the phenomena are described by the expectation values of the observables, i.e., one or more topological coefficients defined at the node level.
2. Statistical Level-2 concerns the computation of the quantities related to probability distribution functions of the above mentioned topological indicators computed at the node level. From this analysis it is possible to characterize power-law shapes in the distribution of the degree of the nodes or in any other indicator. Level-2 analysis allowed for the characterization of some important network regimes and mechanisms as, for instance, the scale-free regime and its associated preferential attachment mechanism.
3. Statistical Level-3 concerns the calculation of the probability measures on the path space, i.e., the cylinder measures on the empirical process. At this level, a phenomenon is described by the probability of a certain configuration of the network that represents the empirical process. Level-3 analysis allowed for the characterization of communities within the network structure.

Level-1 and level-2 analysis are the most common ones and their statistical indicators the most commonly quoted when a stochastic process is analyzed. However to the same expectation values for the observables or to the same pdf's, different processes may be associated. Therefore full understanding of the process requires the determination of the level-3 indicators.

It has been shown (Vilela Mendes et al. (2002), among many others) that the analysis and the reconstruction of a process involves two different but related steps.

- the first step is the identification of the *grammar* of the process, that is, the allowed transitions in the state space or the subspace in path space that corresponds to actual orbits of the system.
- the second step is the identification of the *measure*, which concerns the occurrence frequency of each orbit in typical samples.

Although largely independent from each other, this two features have a related effect on the constraints they impose on the statistical indicators.

In the social sciences and particularly in Economics and Finance, the application of such mathematical devices gave place to the description of a set of empirical findings which are usually called *stylized facts*.

3.3 Stylized Facts in Network Analysis

The notion of stylized facts was once introduced by Nicholas Kaldor (Kaldor 1956) and used thereafter as an encapsulation of regularities found in empirical researches of economic processes. In performing network analysis by means of applying statistical tools developed for the characterization of stochastic process, the concept of

a stylized fact is adequate to identify the empirical evidence of recursive events and relations found in the description of real networks.

In Economics and Finance, some stylized facts use to be formalized as distributions describable by power laws. Others rely just on cross correlation values between stock returns. The main variable that is used to construct the statistical indicators is the differences of log-prices.

$$r(t,n) = \log p\,(t+n) - \log p\,(t) \tag{1}$$

sometimes called the $n-$days return. From the return series of different stocks, there are the following stylized facts:

1. cross correlations between stock returns,
2. non-linear dependencies in the trajectories through time (memory) of a stock return,
3. volatility memory, i.e., memory in the second moment of the volatility of the return of a stock.

Regardless the field of application, some stylized facts have been built on network properties or on the topological coefficients that characterize some network regimes. Most frequent examples are:

1. A power law signature of the distribution of the network degree (as in scale-free networks) are used to characterize heterogeneity, a notion that is used to identify the following different phenomena depending on the field of application, as for instance:
   - the Zipf law in linguistic networks,
   - heterogeneity in word free association in linguistic networks (Solé et al. (2010)),
   - systemic risk in financial networks ((Battiston et al. (2010) and Acemoglu et al. (2013)),
   - contagion phenomenon in epidemic networks (Newman et al. (2003)).
2. (Dis)Assortativeness of node degrees:
   - disassortative mixing in syntactic dependency networks (Ferrer i Cancho et al. (2004)),
   - assortative mixing in semantic networks (Ferrer i Cancho et al. (2004)),
   - capital flows in financial networks (Spelta and Araújo (2012)).
3. Phase transitions associated to characteristic values of topological coefficients:
   - phase transitions in early language acquisition in syntax networks (Solé et al. (2010)),
   - phase transition and systemic risk in financial networks (Acemoglu et al. (2013)).

4. Preferential attachment as the underlying mechanism of network growth:

   - preferential attachment in the evolution of scientific collaboration networks (Newman (2004)),
   - preferential attachment in financial networks (Podobnik et al. (2011)).

Here, some of these stylized facts are used to exemplify the three different levels of statistical analysis and more specifically, the different statistical characterizations that are performed in the analysis of linguistic networks. Several of these examples are also shown in Table 2 (Sect. 5).

3.4 Levels in the Statistical Analysis of Networks

We now look at a set of well-known network measures from the point of view of the statistical levels. Global network indicators typically map certain network properties onto a single value and correspond therefore to the first level (density, clustering coefficient, diameter, etc.). It is clear that a level-1 characterization is relatively rough because many different networks may give rise to the same measure. For instance, a random graph, a regular lattice as well as a scale-free network may well result in the same network density (average degree).

One possibility to differentiate between those cases is to compute a whole set of global level-1 indicators. An example of such a "multi-dimensional" characterization is the so-called small-world regime which is settled on the simultaneous observation, for the whole network, of a low value of the characteristic path length and a high clustering coefficient.

Another way to differentiate between networks that share the same level-1 characteristics is to include the second level of analysis. For instance, the differentiation between a random, a regular and a scale-free graph will be straightforward on the basis of the entire degree distribution. Other measures that are computed at the level of nodes (we may refer to $\mathscr{X}$ as node space) include degree, local clustering and various measures of node centrality.

The second aspect that this section aims to address is to point out that the computation of many of these indicators are in fact based on the third level of analysis. The network diameter (see below), for instance, defined as the longest path in the set of shortest paths between all node pairs in a network, requires the computation of all shortest paths between all node pairs which is related to the path space $\mathscr{X}^{\mathbb{Z}}$.

3.4.1 Node Degrees

Among the main variables that are used to characterize a network is the degree (k_i) of the network nodes (i). For each experimental sample (each network), two main statistical indicators are often computed: first, the average (over i) of k_i, and second

a power law exponent (if it exists) which characterizes the shape of the distribution of $P(k)$. The average degree, or density, is computed as

$$k = \langle k_i \rangle = \frac{1}{n} \sum_{1}^{n} k_i \tag{2}$$

with $\langle \quad \rangle$ meaning the sample average and yielding the average degree (k) of the network.

Second, the degree distribution $P(k)$ of a network is defined as the fraction of nodes in the network with degree k. When the network has n nodes and n_k of them have degree k, then $P(k) = \frac{n_k}{n}$. In case the distribution $P(k)$ follows a power law

$$P(k) = k^{-\lambda} \tag{3}$$

the shape of the distribution is often characterized by a constant λ (typically in the range $2 < \lambda < 3$) and the network is called scale-free.

While network density is probably the most typical level-1 indicator, the exponent λ characterizes the distribution of node degrees and is therefore (by Eq. (3)) related to the second level. Notice however that the computation of both k and λ involve the evaluation of all node degrees, that is $P(k)$. Notice also that many networks may give rise to a certain degree distribution and that the set of networks with the same average degree is even larger. As an additional degree characteristic one may therefore look at the assortativity of node degrees which assesses the degree correlation patterns between pairs of nodes. Despite the fact that assortativity characterizes the network by a single-value and should therefore be considered a level-1 indicator, it assesses characteristics of node pairs, that is, on the space $\mathscr{X}^2$.

3.4.2 Clustering Coefficient

The clustering coefficient is another typical example which is defined at the first and the second statistical level. Namely, in the global variant a single mean clustering value is used as a characterization of the network and this value contains no information about the contributions of individual nodes. On the other hand, the global clustering coefficient is obtained on the basis of a local clustering coefficient defined at the node level $\mathscr{X}$. Consequently, the computation of the global clustering coefficient, while being a first-level indicator, involves the computation of the distribution of the local clustering coefficient.

Notice, however, that the computation of clustering (local and therefore global) involves the computation of the relative frequency of triangles in the network. That is, it involves statistics at the level of triplets of nodes $\mathscr{X}^3$. In Table 1 we denote this as $\mathscr{X}^3 \to \mathscr{X} \to \mathbb{R}$ in order to make clear that the global clustering ($\mathbb{R}$) is based on local clustering ($\mathscr{X}$) which is computed at the as a statistic on node triplets ($\mathscr{X}^3$).

3.4.3 Average Path Length and Diameter

Two other rather typical global indicators (level-1 statistical measures) are the average path length and the network diameter. Both of them are based on network geodesics, that is on the computation of shortest paths between pairs of nodes. While the average path length informs about the average number of steps required to go from one node to another, the diameter is the longest of those. It is clear that both measures are based on the assumption that network trajectories follow shortest paths. It is also clear that both measures map from the third level ($\mathscr{X}^{\mathbb{Z}}$) to the first ($\mathbb{R}$ or respectively $\mathbb{N}$). There is also a node property associated to shortest paths, namely, eccentricity. See Table 1.

3.4.4 Centrality Measures

Centrality measures are a probably the most typical cases of per-node statistics (level-2). They are usually considered as a measure of importance of a node in the network. Various different measures have been proposed, such as degree, betweenness, closeness and eigenvector centrality. All of these measures assess a nodes position in the network with respect to a set of trajectories between pairs of nodes in the graph (see Borgatti (2005) and Borgatti and Everett (2006)). For instance, betweenness quantifies the number of shortest paths in the networks that traverse a given node. Eigenvector centrality, to make another example, is more related to random walks on the network and the respective stationary probability of traversal. Therefore, all of these measures (apart from degrees) define a mapping from the path space of a network to node properties ($\mathscr{X}^{\mathbb{Z}} \to \mathscr{X}$, see Table 1).

4 On the Intelligibility of Statistical Indicators in Linguistic Networks

4.1 Path-Based Measures

Global statistical indicators in network theory typically try to map certain network characteristics onto a single value in order to point out different network regimes with respect to the property at question. However, as shown above, the computation of most of these indicators involves the evaluation of statistics on the path space $\mathscr{X}^{\mathbb{Z}}$. For an overview, see Table 1 below. In this section, we follow the analysis of centrality measures and network flow due to Borgatti (2005) and suggest that, for the interpretation of such measures, it is important to be aware of the implicit assumptions that certain indicators make on network trajectories. Above all in the setting of linguistic networks.

Table 1 High-level network characteristics are mapped onto low-level network indicators

Measure	
density	$\mathscr{X} \to \mathbb{R}$
exp. degree dist.	$\mathscr{X} \dashrightarrow \mathbb{R}$
assortativity	$\mathscr{X}^2 \to \mathscr{X} \to \mathbb{R}$
avg. clustering	$\mathscr{X}^3 \to \mathscr{X} \to \mathbb{R}$
eccentricity	$\mathscr{X}^{\mathbb{Z}} \to \mathscr{X}$
diameter	$\mathscr{X}^{\mathbb{Z}} \to \mathscr{X} \to \mathbb{N}$
centralities	$\mathscr{X}^{\mathbb{Z}} \to \mathscr{X}$
avg. path length	$\mathscr{X}^{\mathbb{Z}} \to \mathscr{X} \to \mathbb{R}$

4.2 Links and Flows, Structure and Function

One of the main contributions of network science in the different areas is that it helps understanding the relation between the structure and the function of a system. For instance, it is now well-known that the connectivity patterns in functional brain networks of Schizophrenia or Alzheimer patients differ in important ways from the patterns in healthy people (e.g., Supekar et al. (2008), Bassett et al. (2009), He et al. (2012), and Zhao et al. (2012)). Moreover, and very importantly, the observed differences like the lack of small-worldness, or clustering enable plausible interpretations about the respective dysfunctions in the brain and we gain, in this way, more insight about the general functioning of that system. Likewise, in traffic networks, power grid networks or networks of inter-bank money transfer the network perspective allows to study the susceptibility to system failure in dependence of certain changes (such as removal of links or nodes) in the structure of the system. Various measures of vulnerability, robustness and stability have been proposed the consideration of which may have very important implications for the design of new, more stable infrastructures.

A better understanding of the relation between the structure of the various linguistic networks and the functioning of the linguistic system must also be at the heart of a unified and applicable theory of linguistic networks. However, as opposed to systems like inter-bank money transfer calling for a type-1 network representation with clear interpretations, in linguistic networks even the interpretation of the functions or processes related to or mapped onto the single connections is not always straightforward, and differs moreover from network type to network type. It is then even more difficult to think in terms of processes taking place in the network as a whole which are typically related to the functioning of the system.

The previous considerations have shown that the majority of statistical network indicators involve computations on the path space of the network and the applicability of these measures is challenged if the question of what flows through the network is undecidable. Borgatti (Borgatti 2005) has discussed these issues at the example of different centrality measures by relating the type of network flow implicit in the

computation of the different measures to the flow in the real system to which those measures have been applied.

4.3 Types of Network Flow

In Borgatti (2005), Borgatti develops a typology of network flows based on two dimensions. The first one relates to different kinds of dynamical processes that flow along the links of a network. Borgatti (2005) proposed that, according to the trajectory, dynamical processes comprise the following types of flows:

1. Geodesics: shortest path to a target destination
2. Paths: no repetition of nodes or links
3. Trails: no repetition of links
4. Walks: no restriction

As a second dimension in this typology, Borgatti (2005) considers the mechanism of node-to-node transmission. The author differentiates:

1. parallel duplication
2. serial duplication, and
3. transfer.

The first one refers to a parallel copying mechanism as present, for instance, in news broad cast. Serial duplication, or copying, refers to the dyadic replication mode in which the information is passed in a serial manner from one node to only one other. As opposed to copying, where the sender does not "loose" the information passed to other nodes, transfer refers to processes in which some thing is transferred, that is, given away, from the sender to the receiver. For Borgatti, the purpose of considering these different kinds of flows is to match different measures of centrality to the different kinds of flows. It turns out that "the most commonly used centrality measures are not appropriate for most of the flows we are routinely interested in" (Borgatti (2005):55).

4.4 Flow in Linguistic Networks

As said, the specification of flow in linguistic networks is not generally straightforward. In networks of semantic relatedness, especially those obtained by word association experiments (Nelson et al. 2004), one could argument on the basis of cognitive processes by which concepts are linked even if a precise understanding of these processes is still lacking. In co-occurrence or precedence networks the situation becomes rather intricate, because it is, in fact, a flow of words that is used to construct a network representation. It is then true that "Paths on network (c) [reference to the co–occurence network] can be understood as the potential universe of

sentences that can be constructed with the given lexicon" (Solé et al. (2010):21), because a text or any other verbal utterance is in fact a sequence of words and therefore naturally defines a trajectory on a word network. On the other hand, the set of trajectories on a co-occurrence network certainly contains a lot of "sentences" that are grammatically incorrect or do not make any sense.

Let us illustrate issues related to the interpretation of network flow with the drastic but simple example of precedence networks. We may conceive a text as a sequence of symbols $S = s_0 s_1 s_2 s_3 \dots s_N$, where, depending on the problem of interest, the symbols s_i correspond to words, lemmata, part-of-speech (PoS) or even to a subset of items such as high-frequent, topic-related nouns. Here we consider words and denote, for convenience, the set of words as $s_i \in \{A, B, C, \dots\}$. In a precedence network (or flow network) an arc from word A to B exists whenever a word pair $(s_i s_{i+1}) = (AB)$ is observed in the text S. In particular, we may put a weight on the arcs corresponding to the frequency with which the associated word pair occurs. In that case, the weighted adjacency matrix (say P) encodes the frequency of all word pairs that are observed in S.

Notice that the degree in such a network representation corresponds to the word frequency and the power law degree distribution is in essence due to the Zipf law. It is noteworthy, moreover, that such a clear interpretation of the network degree in terms of word use statistics is possible because the computation of degrees is based only on node characteristics ($\mathscr{X}$) and not on the network paths.

What kind of network flow (according to Borgatti (2005) in terms of geodesics, paths, trails and walks) can be associated to such a network? Clearly, sentences are not forced to follow shortest paths from a source to a target; a thinking in terms of sources and targets seems to be misplaced in that context. Also the repetition of nodes (words) and links (word pairs) is clearly possible and in fact rather likely. This would mean that sentences, seen as trajectories on word networks, are, in the setting of Borgatti (2005), best classified as walks.

Accordingly, we could normalize the frequency matrix P appropriately in order to contain the probabilities for all words to be followed by the other words, such that P defines a Markov process on $\{A, B, C, \dots\}$. Clearly, such a "model" would be capable of generating all the sentences that have been in the original data S. However, it would generate much more. While it might occasionally generate sentences that make sense and are grammatically correct, this is clearly not the general case. The reason is that the precedence network representation (and co-occurrence more generally) is constructed without sensitivity to grammatical and semantic constraints that are at work in the construction of real sentences. All linguistic relations beyond those to the next word, are just not captured by a precedence network.

In fact, the grammatical constraints in the construction of real sentences which impose some (more complex) restrictions onto the trajectories are not accounted for by the typology proposed in Borgatti (2005). They cannot be accounted for because the information that is needed is just not available in a network representation and its incorporation requires further analysis of the original sequence S. If Borgatti (2005) shows that "centrality measures are not appropriate for most of the flows", the difficulty of interpreting linguistic network flow altogether challenges the

appropriateness of centrality and other path-based measures and their interpretation in the case of linguistic networks.

5 Examples

Table 2 helps to illustrate with some examples the application of the framework for network construction and analysis. Although examples include different fields of application, the purpose of such a framework has a special focus on linguistic networks. The first column indicate the phenomenon at hand and the second column gives an example of the occurrence of the phenomenon. As earlier presented, both network induction and network analysis are performed at three different levels, these correspond to the third and fourth columns, respectively. We recall that while in network design leveling concerns three abstraction levels, in network analysis the three different levels are grounded on the statistical indicators used at each level. The last column in Table 2 provides the main bibliographic reference where the example was reported.

6 Discussion

We classify networks according their level of abstraction and network measures according to the statistical level on which they map. Obviously, the classification of different types of networks according to the level of abstraction they involve is not always clear-cut. Our aim here is not to present such a rigorous classification, but we rather aim at a heuristic perspective to show in what sense linguistic networks are particular. Clearly, in type-1 networks it is usually immediately clear which elements of the real system are encoded into a network description. For instance, speaking of "air traffic network" or "network of inter-bank money transfer" keeps very little space for interpretation: the first one maps air traffic onto a network the second money that is exchanged between banks. Of course, even in systems which call for a network representation (type-1) there is a relative amount of freedom in the design of such a representation. For instance, one could argue that there are many dimensions of (say) air traffic that could be considered, such as passengers, goods, or the number of flights, but it is always clear that a link indicates that something is transfered from one airport to the other, that something flows through the net.

Similarly, for type-2 the meaning of the links is very clear: they indicate the similarity of entities (nodes) with respect to certain features. However, the second type is different from the first especially because the links – encoding similarity – do not generally represent something that really happens between the entities. This is what similarity networks share with type-3 networks.

Indeed, going to the third abstraction level is not such a big step as the proposed classification might indicate. Words linked in a co-occurrence network, for instance, can be seen as words that appear in a similar context, and in that sense, co-occurrence is a kind of similarity. Co-occurrence networks are also nice examples to illustrate that the border between type-3 and type-1 is not always crystal

Table 2 Statistical and abstraction level for several examples of linguistic networks and networks in other fields of application of network theory. In some cases, the classification is not straightforward, see Discussion.

Phenomenon	Example	Sta	Abs	Ref
Small-World	USA power grid	1	1	Watts and Strogatz (1998)
Small-World	Film actors	1	1 or 3	Watts and Strogatz (1998)
Small-World	Air Traffic	1	1	Li and Cai (2004)
Diameter	WorldWideWeb links	1	2	Albert et al. (1999)
Degree, Clustering	Synctatic, Semantic and Co-occurrence	1	3	Solé et al. (2010)
Heterogeneity	Synctatic, Semantic and Co-occurrence	2	3	Solé et al. (2010)
Disassortative mixing	Syntatic nets	2	3	Ferrer i Cancho et al. (2004)
Heterogeneity	Syntatic nets	2	3	Ferrer i Cancho et al. (2004)
Heterogeneity	VWoolf co-occurrence nets	2	3	Ferrer i Cancho et al. (2004)
Degree	Vectorial Semantics	1	2	Salton et al. (1975)
TradeOff Centralities	Airport nets	1	1	Borgatti (2005)
Betweenness	Marriage in Renaissance Florentine	1	1	Borgatti (2005)
Clustering	Financial nets	1	2	Mantegna (1999)
Heterogeneity	Cross-border Debts	2	2	Spelta and Araújo (2012)
Density	Capital flow nets	1	1	Spelta and Araújo (2012)
Heterogeneity	Word free association	2	1	Borge-Holthoefer and Arenas (2010)
Systemic risk	Banking nets	2	1	Battiston et al. (2010)
Systemic risk	Board of Directors	2	3 (bi-partite)	Battiston et al. (2010)
Systemic risk	NYSE network	2	2	Battiston et al. (2010)
Communities	SFI scientists	3	1 or 3	Fortunato and Barthelemy (2007)

clear. On the one hand, we would consider word co-occurrence networks as type-3, corresponding to a high level of abstraction, due to the fact that a link between two words is not obviously related to a *process* in between the two words. As a matter of fact, there are several possibilities of defining a co-occurrence network depending on, for instance, the window size taken into account for the co-occurrence test, and words linked in one co-occurrence design may not be linked in another. On the other hand, however, there are co-occurrence networks in other domains – such as co-authorship – which we might see as networks of a lower abstraction level. In co-authorship networks, two scientists are linked if they are authors of the same article. Even if the specific forms of interchange between the authors is in most cases not visible, it is completely reasonable to consider that a link between the authors maps their conjoint activity and information exchange.

Concerning the statistical levels our main objective is to show that the three different characterizations of a phenomenon at question represent successively finer levels of description of the statistical properties of a network. The second aim is to make clear to which level the different measures relate, because this is important for whether there is a direct interpretation especially in the case of linguistic networks.

We admit that the example of precedence networks, for which we discuss these issues with some detail, is a very special construction and that the argumentation presented in Sect. 4 does not directly apply to all linguistic networks. We assume, however, that the larger class of co-occurrence networks suffers from similar problems concerning the linguistic interpretability of measures that characterize the path space $\mathscr{X}^{\mathbb{Z}}$, because the relation between paths on networks and real word sequences (sentences) is unclear. This may be different for networks of semantic similarity.

In many areas, network representations have proven to be a useful explanatory device which can help to gain insight into the patterns of mutual interdependencies characteristic of complex systems. The flexibility of networks and their applicability to very different phenomena is one of the main reasons for their success and we believe that they are a useful metaphor even if the entities represented by it rely on indirect and observational evidence about the system at question. Accordingly, using networks as an abstraction of linguistic patterns, as a way to map and visualize dependencies between linguistic items may be appropriate and reasonable. In spite of that, however, the transfer of the theory developed for networks to the linguistic field requires a more careful consideration of the context in which this theory has originated and respectively a linguistic assessment of the underlying assumptions.

7 Concluding Remarks

Our contribution relies on highlighting the importance of recognizing network design and network analysis as interdependent tasks. In so doing we developed a framework for network construction and analysis with special focus on linguistic networks. According to such a framework, both network induction and network analysis are performed at three different levels. While in network design leveling concerns three abstraction levels, in network analysis the three different levels are

grounded on the statistical indicators used at each level. Together with the introduction of the framework where network design and network analysis are identified as interdependent tasks, we argue that the level of abstraction at which a given network is induced has an important bearing on this network analysis, particularly on the choice of appropriate topological indicators and on the interpretation of their results. More precisely, we envision that the higher is the abstraction level of network induction, the harder is the interpretation of the topological indicators used in network analysis. We illustrate the framework using examples of linguistic networks as well as some other fields of application of network theory.

These considerations indicate that the field of linguistic networks, by applying well-known statistical tools inspired by network studies in other domains, may, in its current state, have only a limited contribution to the development of linguistic theory. A sophisticated analysis of what topological indicators represent as well as of what they miss is needed in order to advance into that direction. Most importantly, we do not yet have a clear understanding of the trajectories or dynamic processes on the different linguistic networks which makes the use of path-based measures (among them, centrality, average path length, etc.) problematic. On the other hand, the structural differences between linguistic networks of different types (e.g., Pustylnikov (2007)) are clearly indicative of their usefulness in a more applied context as tools for information retrieval and text classification. As soon as we can relate those structural differences to certain linguistic qualities, or show that they represent novel aspects that provide new knowledge of the language system, we may approach to a linguistic network theory. However, it may be that we must go beyond traditional network representation in order to achieve that.

In that regard, we envision the possibility to take the space of linguistic paths seriously in the definition of statistical indicators. Centralities, characteristic path length and other network indicators could be redefined at the level of paths (may be sentences) observed in the original data. Whether this provides new insight and in what sense the resulting measures deviate from their network variant are interesting questions for future research.

Acknowledgements. UECE (Research Unit on Complexity and Economics) is financially supported by FCT (Fundação para a Ciência e a Tecnologia), Portugal. This article is part of the Strategic Project: PEst-OE/EGE/UI0436/2014. SB acknowledges financial support of the German Federal Ministry of Education and Research (BMBF) through the project *Linguistic Networks* (`http://project.linguistic-networks.net`).

References

Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A.: Systemic risk and stability in financial networks. Tech. rep. National Bureau of Economic Research (2013)

Albert, R., Jeong, H., Barabási, A.-L.: Internet: Diameter of the world-wide web. Nature 401(6749), 130–131 (1999)

Arbesman, S., Strogatz, S.H., Vitevitch, M.S.: Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish. Entropy 12(3), 327–337 (2010), doi:10.3390/e12030327

Bassett, D.S., Bullmore, E.T., Meyer-Lindenberg, A., Apud, J.A., Weinberger, D.R., Coppola, R.: Cognitive fitness of cost-efficient brain functional networks. Proceedings of the National Academy of Sciences 106(28), 11747–11752 (2009)

Bassett, D.S., Gazzaniga, M.S.: Understanding complexity in the human brain. Trends in Cognitive Sciences 15(5), 200–209 (2011), doi:10.1016/j.tics.2011.03.006

Battiston, S., Glattfelder, J.B., Garlaschelli, D., Lillo, F., Caldarelli, G.: The structure of financial networks. In: Network Science, pp. 131–163. Springer (2010)

Blanchard, P., Petroni, F., Serva, M., Volchenkov, D.: Geometric representations of language taxonomies. Computer Speech & Language 25(3), 679–699 (2011), doi:10.1016/j.csl.2010.05.003

Borgatti, S.P.: Centrality and network flow. Social Networks 27(1), 55–71 (2005), doi:10.1016/j.socnet.2004.11.008

Borgatti, S.P., Everett, M.G.: A Graph-theoretic perspective on centrality. Social Networks 28(4), 466–484 (2006), doi:http://dx.doi.org/10.1016/j.socnet.2005.11.005

Borge-Holthoefer, J., Arenas, A.: Categorizing words through semantic memory navigation. The European Physical Journal B 74(2), 265–270 (2010) (English), doi:10.1140/epjb/e2010-00058-9

Boss, M., Elsinger, H., Summer, M., Thurner, S.: An empirical analysis of the network structure of the Austrian interbank market. Oesterreichesche Nationalbank Financial stability Report 7, 77–87 (2004)

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer Networks 33(1), 309–320 (2000)

Choudhury, M., Ganguly, N., Maiti, A., Mukherjee, A., Brusch, L., Deutsch, A., Peruani, F.: Modeling discrete combinatorial systems as alphabetic bipartite networks: Theory and applications. Phys. Rev. E 81(3), 036103 (2010), doi:10.1103/PhysRevE.81.036103

Colizza, V., Pastor-Satorras, R., Vespignani, A.: Reaction-diffusion processes and metapopulation models in heterogeneous networks. Nature Physics 3(4), 276–282 (2007)

Ferrer i Cancho, R.: Bibliography on linguistic, cognitive and brain networks (2012), http://www.lsi.upc.edu/~rferrericancho/linguistic_and_cognitive_networks.html

Ferrer i Cancho, R., Mehler, A., Pustylnikov, O., Díaz-Guilera, A.: Correlations in the organization of large-scale syntactic dependency networks. In: TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp. 65–72 (2007)

Ferrer i Cancho, R., Solé, R.V., Köhler, R.: Patterns in syntactic dependency networks. Phys. Rev. E 69(5), 051915 (2004), doi:10.1103/PhysRevE.69.051915

Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1), 36–41 (2007)

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N.V., Corominas-Murtra, B., Bejarano, B., Ardanza-Trevijano, S., Peraita, H., Wall, D.P., Villoslada, P.: The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. Cognitive Processing 12(2), 183–196 (2011) (English), doi:10.1007/s10339-010-0372-x

Grabska-Gradzinska, I., Kulig, A., Kwapien, J., Drozdz, S.: Complex network analysis of literary and scientific texts. International Journal of Modern Physics C 23(07), 1250051 (2012), doi:10.1142/S0129183112500519

Gravino, P., Servedio, V.D.P., Barrat, A., Loreto, V.: Complex Structures and Semantics in Free Word Association. Advances in Complex Systems 15(03n04), 1250054 (2012), doi:10.1142/S0219525912500543

He, H., Sui, J., Yu, Q., Turner, J.A., Ho, B.-C., Sponheim, S.R., Manoach, D.S., Clark, V.P., Calhoun, V.D.: Altered small-world brain networks in schizophrenia patients during working memory performance. PloS One 7(6), e38195 (2012)

Huizinga, H., Nicodème, G.: Are international deposits tax-driven. Journal of Public Economics 88(6), 1093–1118 (2004)

Iyengar, S.R., Veni Madhavan, C.E., Zweig, K.A., Natarajan, A.: Understanding Human Navigation Using Network Analysis. Topics in Cognitive Science 4(1), 121–134 (2012), doi:10.1111/j.1756-8765.2011.01178.x

Kaldor, N.: A model of economic growth. The Economic Journal 67(268), 591–624 (1956)

Konect: OpenFlights network dataset and US airports network dataset – KONECT (2014)

Landauer, T.K., Dutnais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 211–240 (1997)

Lerner, A., Ogrocki, P.K., Thomas, P.J.: Network Graph Analysis of Category Fluency Testing. Cognitive & Behavioral Neurology 22(1), 45–52 (2009)

Li, W., Cai, X.: Statistical analysis of airport network of China. Physical Review E 69(4), 046106 (2004)

Mantegna, R.N.: Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems 11(1), 193–197 (1999)

McGuire, P., Tarashev, N.: Tracking international bank flows. BIS Quarterly Review, 27–40 (2006)

Mehler, A., Geibel, P., Pustylnikov, O.: Structural classifiers of text types: Towards a novel model of text representation. LDV Forum: Zeitschrift für Computerlinguistik und Sprachtechnologie; GLDV-Journal for Computational Linguistics and Language Technology 22(2), 51–66 (2007)

Menon, V.: Large-scale brain networks and psychopathology: a unifying triple network model. Trends in Cognitive Sciences 15(10), 483–506 (2011)

Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: Graph Structure in the Web – Revisited. Accepted paper at the 23rd International World Wide Web Conference (WWW 2014), Web Science Track, Seoul, Korea (April 2014)

Miller, B.A., Bliss, N.T., Wolfe, P.J.: Toward signal processing theory for graphs and non-Euclidean data. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5414–5417. IEEE (2010)

Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)

Minoiu, C., Reyes, J.A.: Network Analysis of Global Banking: 1978-2009. International Monetary Fund (2011)

Mukherjee, A., Choudhury, M., Basu, A., Ganguly, N.: Self-organization of the Sound Inventories: Analysis and Synthesis of the Occurrence and Co-occurrence Networks of Consonants. Journal of Quantitative Linguistics 16(2), 157–184 (2009), doi:10.1080/09296170902734222

Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: The University of South Florida free association, rhyme, and word fragment norms. Behav. Res. Methods. Instrum. Comput. 36(3), 402–407 (2004)

Newman, M.E.J., Barabási, A.-L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2003)

Newman, M.E.J.: Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems 38(2), 321–330 (2004)

Opsahl, T., Agneessens, F., Skvoretz, J.: Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. Social Networks 3(32), 245–251 (2010)

Peng, R.D., Hengartner, N.: Quantitative Analysis of Literary Styles. The American Statistician 56, 2002 (2002)

Podobnik, B., Valentinčič, A., Horvatić, D., Eugene Stanley, H.: Asymmetric Levy flight in financial ratios. Proceedings of the National Academy of Sciences 108(44), 17883–17888 (2011)

Pustylnikov, O.: Guessing Text Type by Structure. In: Proceedings of the 12th ESSLLI Student Session (2007)

Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975), doi:10.1145/361219.361220

Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)

Serva, M., Petroni, F., Volchenkov, D., Wichmann, S.: Malagasy Dialects and the Peopling of Madagascar. J. R. Soc. Interface 9(66), 54–67 (2011)

Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Processing Magazine 30(3), 83–98 (2013)

Soares, M.M., Corso, G., Lucena, L.S.: The network of syllables in Portuguese. Physica A: Statistical Mechanics and its Applications 355(2-4), 678–684 (2005), doi:10.1016/j.physa.2005.03.017

Solé, R.V., Corominas-Murtra, B., Valverde, S., Steels, L.: Language networks: Their structure, function, and evolution. Complexity 15, 20–26 (2010), doi:10.1002/cplx.20305

Soramäki, K., Bech, M.L., Arnold, J., Glass, R.J., Beyeler, W.E.: The topology of interbank payment flows. Physica A: Statistical Mechanics and its Applications 379(1), 317–333 (2007)

Sowa, J.F.: Principles of Semantic Networks. Morgan Kaufmann (1991)

Spelta, A., Araújo, T.: The topology of cross-border exposures:beyond the minimal spanning tree approach. Physica A: Statistical Mechanics and its Applications 391, 5572–5583 (2012)

Sporns, O., Tononi, G., Kötter, R.: The Human Connectome: A Structural Description of the Human Brain. PLoS Comput. Biol. 1(4), e42 (2005), doi:10.1371/journal.pcbi.0010042

Storm, C.: The Semantic Structure of Animal Terms: A Developmental Study. International Journal of Behavioral Development 3(4), 381–407 (1980), doi:10.1177/016502548000300403

Supekar, K., Menon, V., Rubin, D., Musen, M., Greicius, M.D.: Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. PLoS Computational Biology 4(6), e1000100 (2008)

Vilela Mendes, R., Lima, R., Araújo, T.: A process-reconstruction analysis of market fluctuations. International Journal of Theoretical and Applied Finance 5(08), 797–821 (2002)

Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393, 440–442 (1998)

Yu, S., Liu, H., Xu, C.: Statistical properties of Chinese phonemic networks. Physica A: Statistical Mechanics and its Applications 390(7), 1370–1380 (2011), doi:10.1016/j.physa.2010.12.019

Zamora-López, G., Russo, E., Gleiser, P.M., Zhou, C., Kurths, J.: Characterizing the complexity of brain and mind networks. Philosophical Transactions of the Royal Society A 369(1952), 3730–3747 (2011)

Zhao, X., Liu, Y., Wang, X., Liu, B., Xi, Q., Guo, Q., Jiang, H., Jiang, T., Wang, P.: Disrupted small-world brain networks in moderate Alzheimer's disease: a resting-state FMRI study. PloS One 7(3), e33540 (2012)