KONECT – The Koblenz Network Collection

Jérôme Kunegis
Institute for Web Science and Technologies
University of Koblenz–Landau
Koblenz, Germany
kunegis@uni-koblenz.de

ABSTRACT

We present the Koblenz Network Collection (KONECT), a project to collect network datasets in the areas of web science, network science and related areas, as well as provide tools for their analysis. In the cited areas, a surprisingly large number of very heterogeneous data can be modeled as networks and consequently, a unified representation of networks can be used to gain insight into many kinds of problems. Due to the emergence of the World Wide Web in the last decades many such datasets are now openly available. The KONECT project thus has the goal of collecting many diverse network datasets from the Web, and providing a way for their systematic study. The main parts of KONECT are (1) a collection of over 160 network datasets, consisting of directed, undirected, unipartite, bipartite, weighted, unweighted, signed and temporal networks collected from the Web, (2) a Matlab toolbox for network analysis and (3) a website giving a compact overview the various computed statistics and plots. In this paper, we describe KONECT's taxonomy of networks datasets, give an overview of the datasets included, review the supported statistics and plots, and briefly discuss KONECT's role in the area of web science and network science.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement

Keywords

Network analysis, Web observatory

1. INTRODUCTION

Networks are everywhere. Whenever we look at the interactions between things, a network is formed implicitly. In the areas of data mining, machine learning, information retrieval, etc., networks are modeled as *graphs*. Many, if not most problem types can be applied to graphs: clustering, classification, prediction, pattern recognition, and others. Networks arise in almost all areas of research, commerce and daily life in the form of social networks, road networks, communication networks, trust networks, hyperlink networks, chemical interaction networks, neural networks, collaboration networks and lexical networks. The content of text documents is routinely modeled as document–word networks, taste as person–item networks and trust as person–person networks. In recent years, whole database systems have appeared specializing in storing networks. In fact, a majority of research projects in the areas of web mining, web science and related areas uses datasets that can be understood as networks. Unfortunately, results from the literature can often not be compared easily because they use different datasets. What is more, different network datasets have slightly different properties, such as allowing multiple or only single edges between two nodes. In order to provide a unified view on such network datasets, and to allow the application of network analysis methods across disciplines, the KONECT project defines a comprehensive network taxonomy and provides a consistent access to network datasets. To validate this approach on real-world data from the Web, KONECT also provides a large number (>160) of network datasets of different types and different application areas.

KONECT, the Koblenz Network Collection, contains 168 network datasets as of April 2013. In addition to these datasets, KONECT consists of Matlab code to generate statistics and plots about them, which are shown on the KONECT website[1]. KONECT contains networks of all sizes, from small classical datasets from the social sciences such as Kenneth Read's Highland Tribes network with 16 vertices and 58 edges (HT), to the Twitter social network with 52 million nodes and 1.9 billion edges (TF). Figure 1 shows a scatter plot of all networks by the number of nodes and the average degree in the network. Each network in KONECT is represented by a unique two- or three-character code which we write in a sans-serif font, and is indicated in parentheses as used previously in this paragraph. The full list of codes is given online.[2]

This is a short overview paper of KONECT; the full documentation is given on the KONECT website[3], as well as in the upcoming KONECT handbook. This paper first presents KONECT's taxonomy of network datasets in Section 2, then briefly reviews the mathematical background of graph theory in Section 3, then presents a selection of supported network statistics and plots in Sections 4 and 5, and finally outlines the upcoming KONECT Matlab Toolbox in Section 6, as well as the KONECT website in Section 7. We

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1] konect.uni-koblenz.de
[2] konect.uni-koblenz.de/networks
[3] konect.uni-koblenz.de/help

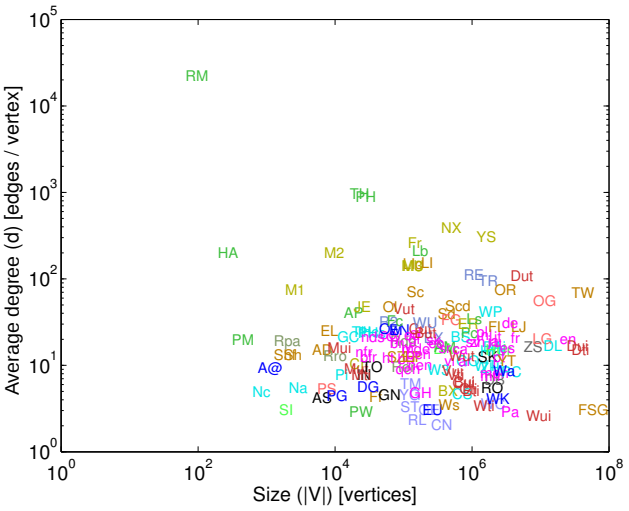


Figure 1: All networks in KONECT arranged by the size (the number of nodes) and the average number of neighbors of all nodes. Each network is represented by a two- or three-character code. The color of each code corresponds to the network category as given in Table 2.

conclude in Section 8 with a short discussion on possible contributions of or to KONECT in the context of the web science community.

2. TAXONOMY OF NETWORKS

Datasets in KONECT represent networks, i.e., a set of nodes connected by links. Networks can be classified by their format (directed/undirected/bipartite), by their edge weight types and multiplicities, by the presence of metadata such as timestamps and node labels, and by the types of objects represented by nodes and links.

The format of a network is always one of the following:

- In **undirected networks** (U), edges are undirected. That is, there is no difference between the edge from u to v and the edge from v to u; both are the edge $\{u, v\}$. An example of an undirected network is the social network of Facebook (Ow), in which there is no difference between the statements "A is a friend of B" and "B is a friend of A."

- In a **directed network** (D), the links are directed. That is, there is a difference between the edge (u, v) and the edge (u, v). Directed networks are sometimes also called *digraphs* (for *directed graphs*), and their edges *arcs*. An example of a directed social network is the follower network of Twitter (TF), in which the fact the user A follows user B does not imply that user B follows user A.

- **Bipartite networks** (B) include two types of nodes, and all edges connect one node type with the other. An example of a bipartite network is a rating graph, consisting of the node types *user* and *movie*, and each rating connects a user and a movie (M3). Bipartite networks are always undirected in KONECT.

Table 1: The edge weight and multiplicity types allowed in KONECT. Each network dataset is exactly of one type.

	Type	Multiple edges	Edge weights
$-$	Unweighted	No	1
$=$	Multiple unweighted	Yes	1
$+$	Positive weights	No	> 0
$\pm$	Signed	No	$\gtrless 0$
$*$	Rating	No	Interval scale
$*^*$	Multiple ratings	Yes	Interval scale

The edge weight and multiplicity types of networks are represented by one of the following six types. The types of edge weights and multiplicities are summarized in Table 1.

- An **unweighted network** ($-$) has edges that are unweighted, and only a single edge is allowed between any two nodes.

- In a **network with multiple edges** ($=$), two nodes can be connected by any number of edges, and all edges are unweighted. This type of network is also called a multigraph.

- In a **positive network** ($+$), edges are annotated with positive weights, and only a single edge is allowed between any node pair.

- In a **signed network** ($\pm$), both positive and negative edges are allowed. Positive and negative edges are represented by positive and negative edge weights. Many networks of this type have only the weights ± 1, but in the general case we allow any nonzero weight.

- **Rating networks** ($*$) have arbitrary real edge weights. They differ from positive and signed networks in that the edge weights are interpreted as an interval scale, and thus the value zero has no special meaning. Adding a constant to all edge weights does not change the semantics of a rating network. Ratings can be discrete, such as the one-to-five star ratings, or continuous, such as a rating given in percent. This type of network allows only a single edge between two nodes.

- **Networks with multiple ratings** ($*^*$) have edges annotated with rating values, and allow multiple edges between two nodes.

Metadata of networks are restricted to timestamps of edges as of this writing, but other metadata of nodes or edges may be added in the future.

- **Temporal networks** (🕒) include a timestamp for each edge, and thus the network can be reconstructed for any moment in the past.

Finally, the network categories classify networks by the type of data they represent. An overview of the categories is given in Table 2.

Affiliation networks are bipartite networks denoting the membership of actors in groups. Groups can be defined as narrowly as individual online communities in which

Table 2: The network categories in KONECT. Each category is assigned a color, which is used in plots, for instance in Figure 1. The property symbols are defined in Table 1. U = Undirected network, D = Directed network, B = Bipartite network. The given dataset counts are current as of April 2013.

Category	Vertices	Edges	Properties	Count
Affiliation	Actors, groups	Memberships	B − =	8
Authorship	Authors, works	Authorships	B − =	18
Co-occurrence	Items	Co-occurrences	U D −	2
Communication	Persons	Messages	U D − =	8
Contact	Persons	Interactions	U D =	4
Features	Items, features	Properties	B −	5
Folksonomy	Users, tags, items	Tag assignments	B =	17
Interaction	Persons, items	Interactions	B − =	14
Lexical	Words	Lexical relationships	U D B − =	5
Physical	Various	Physical connections	U D − =	13
Ratings	Users, items	Ratings	B − $* \ {}_*{}^*$	11
Reference	Documents	References	D − =	28
Semantic	Entities	Relationships	D =	1
Social	Persons	Ties	U D − = + ± ∗	29
Text	Documents, words	Occurrences	B =	5

users have been active (FG) or as broadly as countries (CN). The actors are mainly persons, but can also be other actors such as musical groups. Note that in all affiliation networks we consider, each actor can be in more than one group, as otherwise the network cannot be connected.

Authorship networks are unweighted bipartite networks consisting of links between authors and their works. In some authorship networks such as that of scientific literature (Pa), works have typically only few authors, whereas works in other authorship networks may have many authors, as in Wikipedia articles (en).

Co-occurrence networks represent the simultaneous appearance of items. Co-occurrence networks are unipartite and unweighted. An example is the co-purchase network of Amazon (AM) indicating which persons have purchased the same articles. Note that in most cases, such networks can be derived from another network using a two-mode projection. For instance, a user–user co-purchase network can be derived from a user–item purchase network. In cases where the underlying bipartite network is known, we do not include the two-mode network, as its properties can be derived from the corresponding properties of the original network. As an example, the eigenvalues of the two-mode network's adjacency matrix are the squares of the singular values of the bipartite network's biadjacency matrix. Only when the underlying data is unknown as for the Amazon dataset do we include a co-occurrence network.

Communication networks contain edges that represent individual messages between persons. Communication networks are directed and allow multiple edges. Examples of communication networks are those of emails (EN) and those of Facebook messages (Ow). Note that in some instances, edge directions are not known and KONECT can only provide an undirected network.

Contact networks consist of people and interactions between them. Contact networks are unipartite and allow multiple edges, i.e., there can always be multiple interactions between the same two persons. They can be both directed or undirected. Examples are people that meet each other (RM), or scientists that write a paper together (Pc).

Feature networks are bipartite, and denote any kind of feature assigned to entities. Feature networks are unweighted and have edges that are not annotated with edge creation times. Examples are songs and their genres (GE).

Folksonomies consist of tag assignments connecting a user, an item and a tag. For folksonomies, we follow the 3-bipartite projection approach and consider the three possible bipartite networks, i.e., the user–item, user–tag and item–tag networks. This allows us to apply methods for bipartite graphs to hypergraphs, which is not possible otherwise. Items that are tagged in folksonomies include bookmarks (Dui), scientific publications (Cui) and movies (Mui).

Interaction networks are bipartite networks consisting of people and items, where each edge represents an interaction. In interaction networks, we always allow multiple edges between the same person–item pair. Examples are people writing in forums (UF), commenting on movies (Fc) or listening to songs (Ls).

Lexical networks consist of words from natural languages and their relationships. Relationships can be semantic (i.e, related to the meaning of words) such as the synonym relationship (WO), associative such as when two words are associated with each other by people in experiments (EA), or denote co-occurrence, i.e., the fact that two words co-occur in text (SB). Note that lexical co-occurrence networks are explicitly not included in the broader Co-occurrence category.

Physical networks represent physically existing network structures in the broadest sense. This category covers such diverse data as physical computer networks (TO),

transport networks (OF) and biological food networks (FD).

Rating networks consist of assessments given to items by users, weighted by a rating value. Rating networks are bipartite. Networks in which users can rate other users are not included here, but in the Social category instead. If only a single type of rating is possible, for instance the "favorite" relationship, then rating networks are unweighted. Examples of items that are rated are movies (M3), songs (YS), jokes (JE), and even sexual escorts (SX).

Reference networks consist of citations or hyperlinks between various types of documents. Reference networks are directed. Examples are hyperlinks between pages on the World Wide Web (W3), citations between scientific publications (CS), and citations among patents (PC).

Semantic networks are generic networks of entities connected by relationships. Our dataset collection contains a single semantic network, DBpedia (DB), containing data extracted from the English Wikipedia, in which entities are individual lemmas and relationships are inferred from infoboxes.

Social networks represent ties between persons. Certain social networks allow negative edges, which denote enmity, distrust or dislike. Examples are Facebook friendships (FSG), the Twitter follower relationship (TF), and friends and foes on Slashdot (SZ). Note that some social networks can be argued to be rating networks, for instance the user–user rating network of a dating site (LI). These networks are all included in the Social category.

Text networks consist of text documents containing words. They are bipartite and their nodes are documents and words. Each edge represents the occurrence of a word in a document. Document types are for instance newspaper articles (TR) and Wikipedia articles (EX).

Note that the category system of KONECT is in flux. As networks are added to the collection, large categories are split into smaller ones.

We do not include certain kinds of networks that lack a complex structure. This includes networks without a giant connected component, in which most nodes are not reachable from each other, and trees, in which there is only a single path between any two nodes. Note that bipartite relationships extracted from n-to-1 relationships are therefore excluded, as they lead to a disjoint network. For instance, a bipartite person–city network containing *was-born-in* edges would not be included, as each city would form its own component disconnected from the rest of the network. On the other hand, a band–country network where edges denote the country of origin of individual band members is included, as members of a single band can have different countries of origin. In fact the Countries network (CN) is of this form. Another example is a bipartite song–genre network, which would only be included in KONECT when songs can have multiple genres. As an example of the lack of complex structure when only a single genre is allowed, the degree distribution in such a song–genre network is skewed because all song nodes have degree one, the diameter cannot be computed since the network is disconnected, and each connected component trivially has a diameter of two or less.

3. MATHEMATICAL BACKGROUND

Structures as analysed in KONECT are *networks*. Thus, individual datasets in KONECT are *social networks*, *communication networks*, etc. A network contains nodes and links. Nodes are for instance people or items; links are for instance friendships or ratings. Mathematically, a network is represented by a graph, and we may talk about *graphs* containing *vertices* and *edges*, which correspond to networks containing nodes and links. In most cases, we can use these terms interchangeably, but to be precise, the terms *network/node/link* refer to actually existing structures, whereas *graph/vertex/edge* refer to mathematical objects.

3.1 Definitions

Graphs will be denoted as $G = (V, E)$, in which V is the set of vertices, and E is the set of edges [2]. Without loss of generality, we can assume that the vertices V are consecutive natural numbers, i.e.,

$$V = \{1, 2, 3, \ldots, |V|\}.$$

Edges $e \in E$ will be denoted as sets of two vertices, i.e., $e = \{u, v\}$. We say that two vertices are adjacent if they are connected by an edge; this will be written as $u \sim v$. We say that an edge is incident to a vertex if the edge touches the vertex. We also allow loops, i.e., edges of the form $\{u, u\} = \{u\}$.

In directed networks, edges are pairs instead of sets, i.e., $e = (u, v)$. In directed networks, edges are sometimes called *arcs*; in KONECT, we use the term *edge* for them.

In bipartite graphs, we can partition the set of nodes V into two disjoint sets V_1 and V_2, which we will call the left and right set respectively. Although the assignment of a bipartite network's two node types to left and right sides is mathematically arbitrary, it is chosen in KONECT such that the left nodes are *active* and the right nodes are *passive*. For instance, a rating graph with users and items will always have users on the left since they are active in the sense that it is they who give the ratings.

Networks with multiple edges will be written as $G = (V, E)$, where E is a multiset. The degree of nodes in such networks takes into account multiple edges. Thus, the degree does not equal the number of adjacent nodes but the number of incident edges. When E is a multiset, it can contain the edge $\{u, v\}$ multiple times. Mathematically, we may write $\{u, v\}_1$, $\{u, v\}_2$, etc. Note that we will be lax with this notation. In expressions valid for all types of networks, we will use sums such as $\sum_{\{u,v\}\in E}$ and understand that the sum is over all edges.

In positively weighted networks, we define w as the weight function, returning the edge weight when given an edge. In such networks, the weights are not taken into account when computing the degree.

In a signed network, each edge is assigned a signed weight such as $+1$ or -1. In such networks, we define w to be the signed weight function. In the general case, we allow arbitrary nonzero real numbers, representing degrees of positive and negative edges.

In rating networks, we define r to be the rating function, returning the rating value when given an edge. Note that

rating values are interpreted to be invariant under shifts, i.e., adding a real constant to all ratings in the network must not change the semantics of the network. Thus, we will often make use of the mean rating defined as

$$\mu = \frac{1}{|E|} \sum_{e \in E} r(e).$$

For consistency, we also define the edge weight function w for unweighted and rating networks:

$$w(e) = \begin{cases} 1 & \text{when } G \text{ is unweighted} \\ r(e) - \mu & \text{when } G \text{ is a rating network} \end{cases}$$

We also define a weighting function for node pairs, also denoted w. This function takes into account both the weight of edges and edge multiplicities. It is defined as $w(u, v) = 0$ when the nodes u and v are not connected and if they are connected as

$$w(u,v) = \begin{cases} 1 & \text{when } G \text{ is } - \\ |\{k \mid \{u,v\}_k \in E\}| & \text{when } G \text{ is } = \\ w(\{u,v\}) & \text{when } G \text{ is } + \\ w(\{u,v\}) & \text{when } G \text{ is } \pm \\ r(\{u,v\}) - \mu & \text{when } G \text{ is } * \\ \sum_{\{u,v\}_k \in E} [r(\{u,v\}_k) - \mu] & \text{when } G \text{ is } {}_*{}^* \end{cases}$$

In an unweighted graph $G = (V, E)$, the degree of a vertex is the number of neighbors of that node

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

In networks with multiple edges, the degree takes into account multiple edges, and thus to be precise, it equals the number of incident edges and not the number of adjacent vertices.

$$d(u) = \{\{u, v\}_k \in E \mid v \in V\}$$

In directed graphs, the sum is over all of u's neighbors, regardless of the edge orientation. Note that the sum of the degrees of all nodes always equals twice the number of edges, i.e.,

$$\sum_{v \in V} d(u) = 2|E|.$$

We also define the weight of a node, also denoted by the symbol w, as the sum of the absolute weights of incident edges

$$w(u) = \sum_{\{u,v\} \in E} |w(\{u, v\})|.$$

The weight of a node coincides with the degree of a node in unweighted networks and networks with multiple edges.

3.2 Characteristic Matrices

A very useful representation of graphs is using matrices. In fact, a subfield of graph theory, algebraic graph theory, is devoted to this representation [5]. An unweighted graph $G = (V, E)$ can be represented by a $|V|$-by-$|V|$ matrix containing the values 0 and 1, denoting whether a certain edge between two nodes is present. This matrix is called the adjacency matrix of G and is denoted $\mathbf{A}$. Remember that we assume that the vertices are the natural numbers $1, 2, \ldots, |V|$. Then, the entry $\mathbf{A}_{uv}$ equals one when $\{u, v\} \in E$ and zero when not. This makes $\mathbf{A}$ square and symmetric for undirected graphs, and generally asymmetric (but still square) for directed graphs.

For a bipartite graph $G = (V_1 \cup V_2, E)$, the adjacency matrix has the form

$$\mathbf{A} = \begin{bmatrix} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \end{bmatrix}.$$

The matrix $\mathbf{B}$ is a $|V_1|$-by-$|V_2|$ matrix, and thus generally rectangular. $\mathbf{B}$ is called the biadjacency matrix.

In weighted networks, the adjacency matrix takes into account edge weights. In networks with multiple edges, the adjacency matrix takes into account edge multiplicities. Thus, the general definition of the adjacency matrix is given by

$$\mathbf{A}_{uv} = w(u, v).$$

The degree matrix $\mathbf{D}$ is a diagonal $|V|$-by-$|V|$ matrix containing the absolute weights of all nodes, i.e.,

$$\mathbf{D}_{uu} = |w(u)|.$$

Note that we define the degree matrix explicitly to contain node weights instead of degrees, to be consistent with the definition of $\mathbf{A}$.

The normalized adjacency matrix $\mathbf{N}$ is a $|V|$-by-$|V|$ matrix given by

$$\mathbf{N} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}.$$

Finally, the Laplacian matrix $\mathbf{L}$ is a $|V|$-by-$|V|$ matrix defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}.$$

The KONECT Matlab Toolbox also supports the computation of other characteristic matrices, such as the normalized Laplacian matrix $\mathbf{Z} = \mathbf{I} - \mathbf{N} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$, the stochastic adjacency matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$, and other, more exotic matrices.

4. NETWORK STATISTICS

A network statistic is a numerical value that characterizes a network. KONECT supports the computation of many common network statistics, and this section reviews a selection of these. All network statistics can be computed using the KONECT toolbox, and their values are shown for the KONECT datasets on the website[4].

4.1 Basic Network Statistics

These statistics are simple to define and trivial to compute, and they are reported universally in studies about networks.

The **size** of a network is the number of nodes it contains.

$$n = |V|$$

In a bipartite graph, the size can be decomposed as $n = n_1 + n_2$ with $n_1 = |V_1|$ and $n_2 = |V_2|$.

The **volume** of a network equals the number of edges and is defined as

$$e = |E|.$$

The **average degree** is defined as

$$d = \frac{1}{n} \sum_{u \in V} d(u) = \frac{2e}{n}.$$

[4] konect.uni-koblenz.de/statistics

The average degree is sometimes called the *density*; we avoid that term as it is also used for the fill, as defined next.

The **fill** of a network is the proportion of edges to the total number of possible edges.

$$\begin{aligned} f &= 2e/[n(n+1)] && \text{when } G \text{ is undirected} \\ f &= e/n^2 && \text{when } G \text{ is directed} \\ f &= e/(n_1 n_2) && \text{when } G \text{ is bipartite} \end{aligned}$$

In the undirected case, the expression is explained by the fact that the total number of possible edges is $n(n+1)/2$ including loops. The corresponding numbers for directed and bipartite networks are n^2 and $n_1 n_2$.

The **maximum degree** equals the highest degree value attained by any node.

$$d_{\max} = \max_{u \in V} d(u)$$

4.2 Connectivity Statistics

Connectivity statistics measure to what extent a network is connected. Two nodes are said to be connected when they are either directly connected through an edge, or indirectly through a path of several edges. A connected component is a set of vertices all of which are connected, and unconnected to the other nodes in the network. The largest connected component in a network is usually very large and called the giant connected component. When it contains all nodes, the network is connected.

The **size of the largest connected component** is denoted as CC. In bipartite networks, the number of left and right nodes in the largest connected components are denoted CC_1 and CC_2, with $\mathrm{CC}_1 + \mathrm{CC}_2 = \mathrm{CC}$.

The **relative size of the largest connected component** equals the size of the largest connected component divided by the size of the network

$$\mathrm{CC}_{\mathrm{rel}} = \frac{\mathrm{CC}}{n}.$$

In directed networks, we additionally define the **size of the largest strongly connected component** CC_{s}. A strongly connected component is a set of vertices in a directed graph such that any node is reachable from any other node using a path following only directed edges in the forward direction. We always have $\mathrm{CC}_{\mathrm{s}} \leq \mathrm{CC}$.

4.3 Path Length Statistics

These statistics are based on the length of paths in the network.

The **diameter** δ of a graph equals the longest shortest path in the network. Given two vertices in a graph, their distance is defined as the minimal number of edges needed to reach one from the other. The largest distance attained between two nodes is then the diameter. Note that the diameter is undefined (or infinite) in unconnected networks, and thus we consider always the diameter of the network's largest connected component. The diameter can be computed exactly or estimated, in which case it is noted $\tilde{\delta}$.

A related statistic is the **90-percentile effective diameter** $\delta_{0.9}$, which equals the number of edges needed on average to reach 90% of all other nodes.

The **median path length** δ_M is the median length of shortest paths in the network, and the **mean path length** δ_{m} is the mean over the shortest path lengths for all node pairs in the network. Both the median and mean path lengths are computed taking into account node pairs of the form (u, u).

4.4 Degree Distribution Statistics

The distribution of degree values $d(u)$ over all nodes u is often taken to characterize a network. Thus, a certain number of network statistics are based solely on this distribution, regardless of overall network structure.

The **power law exponent** is a number that characterizes the degrees of the nodes in the network. In many circumstances, networks are modeled to follow a degree distribution power law, i.e., the number of nodes with degree n is taken to be proportional to the power $n^{-\gamma}$, for a constant γ larger than one [1]. This constant γ is called the power law exponent. Given a network, its degree distribution can be used to estimate a value γ. There are multiple ways of estimating γ, and thus a network does not have a single definite value if it. In KONECT, we estimate γ using the robust method given in [7, Eq. 5]

$$\gamma = 1 + n \left(\sum_{v \in V} \ln \frac{d(u)}{d_{\min}} \right)^{-1},$$

in which $d_{\min}$ is the minimal degree.

The **Gini coefficient** is a measure of inequality from economics, typically applied to distributions of wealth or income. In KONECT, we apply it to the degree distribution, as described in [6]. The Gini coefficient can either be defined in terms of the Lorenz curve, a type of plot that visualizes the inequality of a distribution, or using the following expression. Let $d_1 \leq d_2 \leq \cdots \leq d_{|V|}$ be the sorted list of degrees in the network. Then, the Gini coefficient is defined as

$$G = \frac{2 \sum_{i=1}^{n} i d_i}{n \sum_{i=1}^{n} d_i} - \frac{n+1}{n}.$$

The Gini coefficient takes values between zero and one, with zero denoting total equality between degrees, and one denoting the dominance of a single node.

4.5 Algebraic Statistics

Algebraic statistics are based on the eigenvalues of a network's characteristic matrices. They are motivated by the broader field of spectral graph theory, which characterizes graphs using the spectra of these matrices [3]. In the following we will denote by $\lambda_k[\mathbf{X}]$ the k^{th} dominant eigenvalue of the matrix $\mathbf{X}$. For the adjacency matrix $\mathbf{A}$, the dominant eigenvalues are the largest absolute ones; for the Laplacian $\mathbf{L}$, they are the smallest ones. Also, the matrix $\mathbf{L}$ will only be considered for the network's largest connected component.

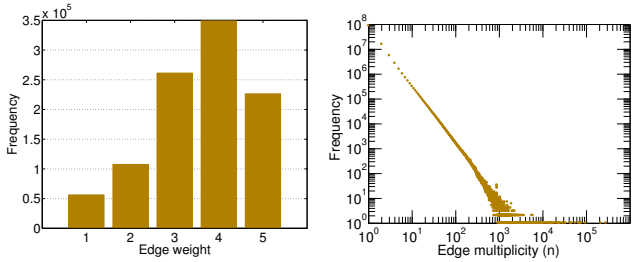
The **spectral norm** of a network equals the spectral norm (i.e., the largest absolute eigenvalue) of the network's adjacency matrix

$$|\lambda_1[\mathbf{A}]| = \|\mathbf{A}\|_2 .$$

The spectral norm can be understood as an alternative measure of the size of a network.

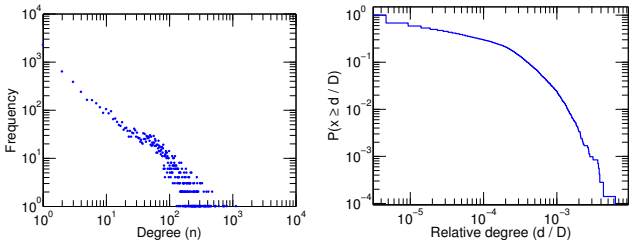
The **algebraic connectivity** equals the second smallest nonzero eigenvalue of $\mathbf{L}$ [4]

$$a = \lambda_2[\mathbf{L}].$$



(a) Edge weight distribution (b) Edge multiplicity distribution

Figure 2: The distribution of (a) edge weights for the MovieLens rating network (M2) and (b) edge multiplicities for the English Wikipedia edit network (en).



(a) Degree distribution (b) Cumulative degree distribution

Figure 3: The degree distribution and cumulative degree distribution for the Wikipedia election network (EL).

The algebraic connectivity is zero when the network is disconnected – this is one reason why we restrict the matrix $\mathbf{L}$ to each network's giant connected component. The algebraic connectivity is larger the better the network's largest connected component is connected.

4.6 Other Statistics

The **clustering coefficient** is a statistic of unipartite networks that measures to what extent two incident edges tend to be completed by a third edge to form a triangle.

$$c = \frac{|\{u, v, w \in V \mid u \sim v \sim w \sim u\}|}{|\{u, v, w \in V \mid u \sim v \sim w\}|}$$

The clustering coefficient has values between zero and one, with a value of one denoting that all possible triangles are formed (i.e., the network consists of disconnected cliques), and zero when it is triangle free. Note that the clustering coefficient is trivially zero for bipartite graphs.

5. PLOTS

Plots are drawn to visualize a certain aspect of a dataset. For instance, a large fraction of studies analysing networks will show a plot of the degree distribution. This plot and many others can be used to compare several network visually, or to illustrate the definition of a certain numerical statistic. In the following, we show a selection of plots supported by KONECT, intended to illustrate the range of supported plots. As a running example, we show the plots for the Wikipedia elections network (EL). Plots for all networks (in which computation was feasible) are shown on the KONECT website[5]. The KONECT Matlab toolbox contains code for generating these plot types.

Edge Weight and Multiplicity Distribution.

The edge weight and multiplicity distribution plots show the distribution of edge weights and of edge multiplicities for positive/signed/rating networks and multiple networks respectively. They are not generated for unweighted networks. The X axis shows values of the edge weights or multiplicities, and the Y axis shows frequencies. Edge multiplicity distributions are plotted on doubly logarithmic scales. Examples of both plots are given in Figure 2.

[5]konect.uni-koblenz.de/plots

Degree Distribution.

The distribution of degree values $d(u)$ over all vertices u characterizes the network as a whole, and is often used to visualize a network. In particular, a power law is often assumed, stating that the number of nodes with n neighbors is proportional to $n^{-\gamma}$, for a constant γ [1]. This assumption can be inspected visually by plotting the degree distribution on a doubly logarithmic scale, on which a power law renders as a straight line. KONECT supports two different plots: The degree distribution, and the cumulative degree distribution. The degree distribution shows the number of nodes with degree n, in function of n. The cumulative degree distribution shows the probability that the degree of a node picked at random is larger than n, in function of n. Both plots use a doubly logarithmic scale. Examples of both plots are given in Figure 3.

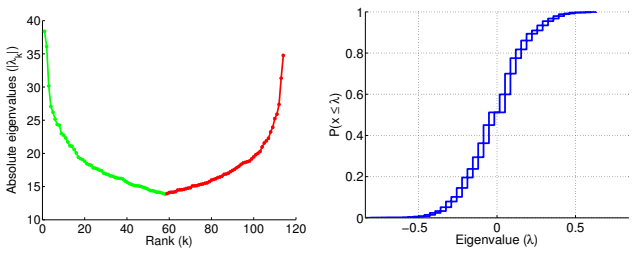
Another visualization of the degree distribution supported by KONECT is in the form of the Lorenz curve, a type of plot to measure inequality originally used in economics (not shown).

Spectral Plots.

The eigenvalues of a network's characteristic matrices $\mathbf{A}$, $\mathbf{N}$ and $\mathbf{L}$ are often used to characterize the network as a whole. KONECT supports computing and visualizing the spectrum (i.e., the set of eigenvalues) of a network in multiple ways. Two types of plots are supported: Those showing the top-k eigenvalues computed exactly, and those showing the overall distribution of eigenvalues, computed approximately. The eigenvalues of $\mathbf{A}$ are positive and negative reals, the eigenvalues of $\mathbf{N}$ are in the range $[-1, +1]$, and the eigenvalues of $\mathbf{L}$ are all nonnegative. For $\mathbf{A}$ and $\mathbf{N}$, the largest absolute eigenvalues are used, while for $\mathbf{L}$ the smallest eigenvalues are used. Examples of the spectra of $\mathbf{A}$ and $\mathbf{N}$ are shown in Figure 4.

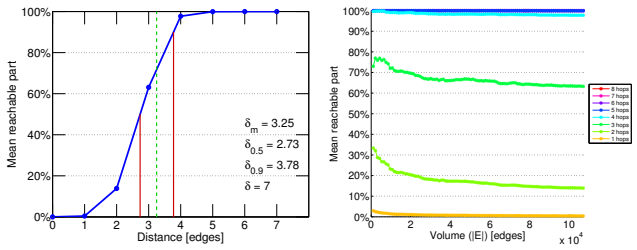
Hop Plot.

Path length statistics can be visualized in the hop plot. The hop plot shows, for each integer k, the number of node pairs at distance k from each other, divided by the total number of node pairs. The hop plot can be used to read off the diameter, the median path length, and the 90-percentile effective diameter (see Section 4.3). For temporal networks, the hop plot can be shown over time. Examples are computed in Figure 5.



(a) Top-k eigenvalues of **A** (b) Cumulative eigenvalue distribution of **N**

Figure 4: The top-k eigenvalues of A and the cumulative spectral distribution of N for the Wikipedia election network (EL). In the first plot (a), positive eigenvalues are shown in green and negative ones in red.



(a) Hop plot (b) Temporal hop plot

Figure 5: The hop plot and temporal hop plot for the Wikipedia election network (EL).

6. MATLAB TOOLBOX

The KONECT Matlab Toolbox[6] is a set of functions for the Matlab programming language[7] containing implementations of statistics, plots and other network analysis methods. The KONECT Matlab Toolbox was used to generate the numerical statistics and plots in this paper as well as on the KONECT website.

7. WEBSITE

The KONECT website, available at konect.uni-koblenz.de contains information about all datasets, all numerical statistics, plots, as well as downloads of all datasets for which it is legal to do so. More than half of all datasets are available for download. The datasets are provided in a unified and simple format (edge-wise tab-separated values) to allow the easy analysis of networks in many programming languages and environments. The website also provides downloads of all datasets for which this is possible within legal constraints, and the extraction code which was used to generate the datasets.

8. DISCUSSION

KONECT is a new entry in the field of network dataset collections, and as such it competes, or rather complements, collections such as Albert-László Barabási's resources[8], Tore Opsahl's collection[9], the Pajek datasets[10], Jure Leskovec's SNAP[11] and many others. What sets KONECT apart from these is (1) its large size (>160 datasets), (2) a comprehensive taxonomy of network datasets, (3) its broad scope of 15+ categories, (4) an integrated Toolbox for computation of both statistics and plot, and (5) the integration of statistics and plots into the website, for easy exploration of the corpus. Rather than seeing these related collections as competitors, we see them as collaborators, and in fact many KONECT datasets are imported from them.

[6]konect.uni-koblenz.de/toolbox
[7]www.mathworks.com/products/matlab
[8]www3.nd.edu/∼networks/resources.htm
[9]toreopsahl.com/datasets
[10]vlado.fmf.uni-lj.si/pub/networks/data
[11]snap.stanford.edu

By disseminating KONECT as part of the Web Observatory Workshop, we hope to raise awareness for the effort this project represents, and in particular solicit the participation of practitioners in the acquisition of new datasets, in the usage of KONECT datasets in studies, and in the shaping of KONECT's future. In particular, we want to receive feedback and stimulate discussions about (1) the prominent role of networks in the context of web science, (2) the importance of using as many datasets as possible in studies, and thus of the importance of dataset repositories such as KONECT, (3) datasets and dataset types that researchers want to use, (4) datasets and dataset types that researchers can contribute, (5) which statistics and plots are most important to users and (6) which data formats would facilitate adoption of KONECT. As an example for the last point, we are currently experimenting with releasing datasets in a Semantic Web-compatible way in RDF format.

9. ACKNOWLEDGMENTS

Datasets in KONECT have been extracted and prepared by Jérôme Kunegis, Martina Sekulla, Daniel Dünker and Holger Heinz. We thank everyone who has made network dataset available publicly. The research leading to these results has received funding from the European Community's Seventh Frame Programme under grant agreement n° 257859, ROBUST.

10. REFERENCES

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
[2] B. Bollobás. *Modern Graph Theory*. Springer, 1998.
[3] F. Chung. *Spectral Graph Theory*. American Math. Soc., 1997.
[4] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Math. J.*, 23(98):298–305, 1973.
[5] C. D. Godsil and G. Royle. *Algebraic Graph Theory*. Springer, 2001.
[6] J. Kunegis and J. Preusse. Fairness on the web: Alternatives to the power law. In *Proc. Web Science Conf.*, pages 175–184, 2012.
[7] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Phys.*, 46(5):323–351, 2006.