S

Sources of Network Data

Monika Cerinšek[1] and Vladimir Batagelj[2,3]
[1]Abelium d.o.o, Ljubljana, Slovenia
[2]Department of Theoretical Computer Science, Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
[3]University of Primorska, Andrej Marušič Institute, Koper, Slovenia

Synonyms

Almost network data; Archives; Boundary problem; Copyrights; Databases; Ego-centered networks; Ethics; Networks; Observation; Random networks; Semantic web; Surveys

Glossary

Cloud Technology	A use of hardware and software that are delivered as a service over a network (usually the Internet)
Computer-Assisted Text Analysis – CATA	Techniques that model and structure the information content of textual sources on a computer
Genealogy	A study of families and tracing of their lineages
Network Analysis	A study of networks as a representation of relations between discrete objects
Social Network	A social structure based on a set of actors (individuals or organizations) and the ties between these actors
Web Crawler	An Internet bot that automatically browses the World Wide Web

Definition

In network data different entities are linked through their relations. They can be found in many forms and obtained from observations, surveys, archives, databases, etc. Network data can also be generated from other types of data, semantic web, or even be randomly generated.

Introduction

We can find the network data almost everywhere in our lives:

- Cities are linked with roads.
- People in a group are linked by exchange of messages (mail, phone).
- Works from a field of research are linked with citations.
- Researchers are linked through their collaborations.

R. Alhajj, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*,
https://doi.org/10.1007/978-1-4614-7163-9_313-1

- Atoms in molecules are linked with their chemical bonds.
- Words are linked according to their co-appearances in sentences of some text.
- In genealogies people are linked by marriage and parent-child ties.

A ***graph*** $\mathcal{G}$ is an ordered pair of sets $(\mathcal{V}, \mathcal{L})$, with the set of ***nodes*** $\mathcal{V}$ and the set of ***links*** $\mathcal{L}$. Every link has two end-nodes. Every link is either directed, an ***arc***, or undirected, an ***edge***. A ***network*** $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W})$ consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, describing the structure of a network, and additional data: ***properties*** $\wp$ of nodes and ***weights*** $\mathcal{W}$ on links.

There are different types of networks in addition to the ordinary networks.

A ***two-mode*** network is a network $\mathcal{N} = ((\mathcal{I}, \mathcal{J}), \mathcal{L}, \wp, \mathcal{W})$, where the set of nodes $\mathcal{V} = \mathcal{I} \bigcup \mathcal{J}$ is split into two disjoint subsets of nodes $\mathcal{I}$ and $\mathcal{J}$ and each link from $\mathcal{L}$ has one end-node in $\mathcal{I}$ and the other end-node in $\mathcal{J}$.

A ***multirelational*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W})$ allows multiple relations to exist in the network $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$.

In a ***temporal*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \wp, \mathcal{W}, \mathcal{T})$ the time $\mathcal{T}$ is attached to a network. For all nodes and links we have to specify the time intervals in which the element is active (present) in the network. Also properties and weights can change through time – we describe their values as temporal quantities (Batagelj and Praprotnik 2016).

Sometimes a given system is described with a ***collection*** of (often two-mode) ***linked*** networks. For example, bibliographic networks (Batagelj and Cerinšek 2013) and MetaMatrix (Carley 2003).

When constructing a network we must first specify what the nodes are and which relation is linking them – the ***network boundary*** problem (Wasserman and Faust 1994; Marsden 2011). According to the plan of network analyses, we need to bound the set of nodes to those that we need. Along with nodes and links, we also select their properties. We have to decide whether the network is one-mode or two-mode and which node properties are important for our intended analyses. We have to answer several questions about the links: Are the links directed? Are there different types of links (relations)? Can a pair of nodes be linked with multiple links? What are the weights on the links? Is the network static, or is it changing through time?

Sometimes the list of nodes is known in advance (e.g., students in the class). But often the set of nodes is constructed during the network data collection process. In this case we have to specify the membership criteria determining for each potential node whether it belongs to the network or not.

For collecting the network data, the ***snowball*** procedure is often used. We first select a (small) set of nodes as initial candidates. Then we collect the data about each candidate and determine its neighboring nodes. The new ones among them we add to the list of candidates. The inclusion of the new nodes can also be determined by some other criteria, for example, by the distance from the closest initial node. We end this process when the list of candidates is exhausted, or the limit to the number of inspected nodes is reached.

Another problem that often occurs when defining the set of nodes is the ***identification*** (entity resolution, disambiguation) of nodes. The unit corresponding to a node can have different names (***synonymy***), or the same name can denote different units (***homonymy*** or ***ambiguity***). For example, in a bibliography on mathematics from Zentralblatt MATH, the names Borštnik, N. S. Mankoč; Mankoč Borštnik, N.; Mankoč-Borštnik, Norma; Mankoč Borštnik, Norma Susana; Mankoc-Borstnik, N.S.; and Mankoč Borštnik, N.S. belong to the same author. On the other hand, in Zentralblatt MATH at least two different Smith, John W. are recorded, because publications of the author(s) with this name spanned from 1868 to 2007. There are at least 623 different mathematicians with the name Zhang, Li in the MathSciNet Database. Its editors are trying hard, from the year 1985, to resolve the author's identification problem (Martin et al. 2013) during the data entry phase. In the future

the problem could be eliminated by general adoption of initiatives such as ResearcherID or ORCID.

The identification problem also appears when the units are extracted from the plain text, for example, "the President of the USA" and "Barack Obama." To resolve it we have to provide lists of equivalent terms. Another source of identification problems is the grammar rules of the language used in text. For example, the action "go" can appear in the text in different forms: "go," "goes," "gone," "going," "went". To resolve these problems we apply the stemming or lemmatization procedures from natural language processing toolkits such as NLTK or MontyLingua.

A special approach of collecting data for a network analysis is by forming ***ego-centered*** networks (Lozar Manfreda et al. 2004). This approach is used when the population of our interest is too large. From the population we select a sample of units (***egos***) and collect the data about them and their neighbors (***alters***) and links among them. An example of ego-centered networks is the friendship networks of selected persons from Facebook.

Collecting the network data we have to respect legal (copyright, privacy) and ethical constraints (Borgatti and Molina 2003; Eynon et al. 2008; Charlesworth 2008; Breiger 2005).

The network data can be obtained in many ways:

- By observation
- With surveys or interviews
- From archives and databases
- From data organized in a network form
- Derived from the data
- From semantic web
- With generating random networks

Each of the above methods for gathering the network data is described in more details in the following subsections. For details additional references are provided.

Observation

To form a network we must first obtain the data. The ways of obtaining the data have been changing through history following the technological developments. A basic approach is the observation (Mitchell 1969). The observation is a human activity consisting of receiving information about the outside world through the senses, or the recording of data using scientific instruments and includes also any data collected during this activity. Scientific instruments were developed to amplify human powers of observation, such as weighing scales, clocks, telescopes, microscopes, thermometers, cameras, and tape recorders, and also to translate into perceptible form events that are unobservable by human senses, such as voltmeters, spectrometers, infrared cameras, oscilloscopes, interferometers, Geiger counters, x-ray machines, and radio receivers (Shipman et al. 2009).

Making direct measurements is the most accurate method for many variables but can be limited by the technology available. The main alternative to direct observation is to require others to report their activities.

An example of the observational network data collection is described in the PhD thesis of Sampson (1968). He did an ethnographic study of community structure in a New England monastery – he divided 18 novices into 4 groups at 5 time points based on his observations and analyses. Another example is the detection of molecular structure of organic molecules.

Surveys

Survey is a data-gathering method that actively includes the observed people (Marsden 1990). They allow us to study attitudes, beliefs, behaviors, and other characteristics. With carefully prepared questionnaires one can collect vast amount of quality data. A ***questionnaire*** is a list of questions. Answers can be ***closed*** – selected from a given list. They are easier to analyze. But the ***open*** answers, that are not given in advance, allow the analysts to get a wider amount of information.

A survey can take different forms: face-to-face, paper and pencil, telephone, e-mail, or online. Nowadays questionnaires are mostly digital (online surveys) that allow them to be adaptable, immediate checking of the entered data, and also collecting some contextual (observational) data.

The use of direct observation in combination with surveys can provide additional information. It can confirm or negate information gained from surveys. As observation itself also the observation in combination with surveys must be prepared. The observant might use appropriate scales, checklists, and other observation materials that are selected in accordance with the questions and possible closed answers in the survey.

An interesting network obtained by interviewing is the *Edinburgh Associative Thesaurus*.

Surveys are the most commonly used methods to gather social network data. They are also used to study interorganizational relations (Mizruchi and Galaskiewicz 1993). For details on surveys and questionnaires, see the entry "► Questionnaires for Measuring Social Network Contacts."

Archives and Databases

An archive is a collection of historical data, or the physical place where they are located (Schmidt 2011). Archives have a historical, cultural, and evidentiary value. Archives exist everywhere, where data has been stored. Every organization has an archive of past activities; universities have archives of past students' achievements and research; backup on a personal computer is an archive of past usage of the computer, etc. With the transition of office work to computers and the spread of Internet, many archives became digital. A database is an organized collection of data, mostly in digital form. Database is organized in records and for each record it has some properties stored (Ullman and Widom 2008). Because data is organized, it is very easy to transform it in a collection of (often two-mode) networks which are then used in the network analysis. Smaller amounts of data can be presented in a tabular form as spreadsheets.

For example, there exist many bibliographic databases (Web of Science, Scopus, Zentralblatt MATH, etc.) that are keeping data about published papers and books. Even the World Wide Web is being partially collected and preserved as an archive for future researchers, historians, and the public.

As a source of data, the archives of various kinds are inexpensive and advantageous for studying especially social networks in the past (Marsden 1990). The network data can be derived from archived data. For example, relations between corporations can be studied based on information about persons on the boards of directors of the corporations.

Historical archives help researchers to gain knowledge about the development of some field – economics, scholar, military, etc. For example, with data from World War II one can study the military movements through the war, the transfer of refugees or prisoners, the transfer of weapons, etc. Another example is the analysis of alliances between the most powerful countries over a selected time period.

Archived data about the inhabitants of a city or an area can be used for genealogical analysis. In genealogy we can search for typical marriage patterns and their irregularities. For example, marriages among relatives to keep the family's wealth, or on the other hand, marriages outside the family to increase its influence. The genealogical data are often available in the GEDCOM format. Large collection of family genealogies is available at the *Genealogy Forum*. For "scientific" genealogies used in anthropological research, see the site *KinSource*.

Activities on the Internet, such as e-mail, chat, and forums, leave traces that can be used as sources for network data. A notorious example is the *Enron e-mail data*.

Especially interesting for network analysis is the World Wide Web as an archive. The web crawlers visit the page with URL from the list of URLs, identify all hyperlinks in it, and add the URLs of these hyperlinks to the list. The largest web archiving organization based on crawling approach is the *Internet Archive*, but also national libraries, national archives, and other

organizations are involved in archiving mostly culturally important web content (*Web Archiving Service*).

Enormous archives are being formed by different social networking services such as Facebook, Twitter, LinkedIn, and Google+. These organizations are collecting the data about users, their posts, or tweets. Data about users are not publicly available. The user can download only the data about his past activity and the data that other users declared visible to him.

A large amount of data is stored in Internet Movie Database (*IMDb*) and services such as Amazon, lastFM, *Pandora*, or *Netflix*. Converting data into multiple two-mode networks and combining them in network analysis allows us to obtain information about collaboration between actors, producers and composers, similarity of the movies according to different measures, etc.

With the development of technology, different types of databases occurred, where the type of the database is defined with the way the data is stored in a database. With growth of available data the data warehouses were developed. Data warehouses archive data directly from the source. It is a central source of data for use by managers for creating statistical dashboards and reports about it. The other very popular type of database is cloud database that relies on the cloud technology (Voorsluys et al. 2011).

A graph database (Angles and Gutierrez 2008) is also useful in the network analysis and it is interesting because of the way the data is stored in it. It uses graph structure to represent and store information. Specialized graph database uses a network model, which is conceived as a flexible way of representing objects and their relationships. See for example the Neo4j support of the Panama Papers (ICIJ 2016).

Every day large amounts of data are being collected. Big data (White 2012) is considered to be a collection of large data sets. These data sets are so large and complex that it is very difficult to process them using traditional data processing applications. Also suitable technologies are required such as cluster analysis, machine learning, neural networks, pattern recognition, and anomaly detection.

Many repositories of networks and datasets of other types are available: *Repositories of Datasets*, *KDnuggets Datasets for Data Mining*, *Data Surfing on the World Wide Web*, *Public Data Sets on Amazon Web Services*, *TunedIT*, the *Internet2 Observatory Data Collections*, *Infochimps*, the Cooperative Association for Internet Data Analysis (*CAIDA Data*), *Network Data Sources on Pajek's web page*, *KONECT*, *SNAP*, *GDELT*, *NYC Taxi and Uber Trips*, *Bike sharing*, and *MAG*.

Different activities are traced by their logs. Mobile network operators record the usage of the phones by their users, the data from weather stations is collected, online social network providers collect the data about their users (Abdesslem et al. 2012), different sensor networks are being established, peer-to-peer (P2P) networks are more and more interesting, using the radio-frequency identification (RFID) tags we can follow the movement of their owners or collect the transportation statistics, etc. Such data can be used for prediction or just for the behavioral analysis of the users.

Almost Network Data

Some data are already organized in a network form. A transportation network is a network of roads, pipes, streets, or any other similar structure that allows transportation of some kind. They are represented as links, and crossings are presented as nodes. Another area that deals with a lot of data in a network form is chemistry. The structure of every molecule is a network with atoms as nodes and covalent chemical bonds as links between them. The most interesting for network analysis are organic molecules such as proteins, lipids, hydrocarbons, and DNA. A lot of chemical and biological data is available at *Ensembl*, *GO Database*, *KEGG*, and *Protein Data Bank*.

To analyze such data using the selected network analysis tool, we usually have to transform them into the corresponding input network data format. These issues are elaborated in details in the entry "► Network Data File Formats."

Sometimes special programming solutions should be developed to perform the required transformation. For example, the transformation of the ESRI shape file describing the map of borders between the country's administrative units (states, counties) into the neighborhood relation of the administrative units can be done with a short program in R using the function `poly2nb` from the package `spdep`.

Networks Derived from Data

Some data sources require more sophisticated procedures to transform them into corresponding networks.

Very intriguing data sources are also the daily news archives of the news agencies (Agence France-Presse, Reuters, United Press International, American Press Agency, Xinhua, ITAR-TASS, etc.). A single news is essentially a (tagged) plain text that can be analyzed with ***computer-assisted text analysis*** (Popping 2000). One of the main approaches to this type of text analysis is the semantic text analysis. The units of the text are encoded according to the Chomsky's ***subject-verb-object*** model which can be directly transformed into temporal multirelational networks with subjects and objects as nodes and verbs as relations. Examples of applications of this approach are the *Kansas Event Data System*, *Paul Hensel's International Relations Data Site*, or *Correlates of War*. An elaboration of this approach is given in the Franzosi's book *From Words to Numbers* (Franzosi 2004). See also the *Centering Resonance Analysis approach* proposed by Steve Corman.

Another example is the neighbors' networks. Let $\mathcal{V}$ be a set of (multivariate) units and $d(u,v)$ a ***dissimilarity*** on it. They determine two types of networks:

the ***k-nearest neighbors*** network: $\mathcal{N}(k) = (\mathcal{V}, \mathcal{A}, w)$.
$(u, v) \in \mathcal{A}$ iff v is among k nearest neighbors of u, $w(u,v) = d(u,v)$.
and the ***r-neighbors*** network: $\mathcal{N}(r) = (\mathcal{V}, \mathcal{E}, w)$.
$(u, v) \in \mathcal{E}$ iff $d(u,v) \leq r$ and $w(u,v) = w(v,u) = d(u,v)$.

These networks provide a link between (multivariate) data analysis and network analysis. For larger sets of units a problem of an efficient algorithm for determining the nearest neighbors arises. David M. Mount wrote the *Approximate Nearest Neighbor Library* with fast algorithms for the (approximate) nearest neighbor search. In R these algorithms are available through the function `ann` in package `yaImpute`.

Semantic Web

Semantic web (Berners-Lee et al. 2001) is an upgrade and an extension of the ordinary web. It provides a data layer in the World Wide Web to be used by web services. The basis for semantic web is the semantic description of the web content with the use of metadata and ontologies. The aim is to convert web of unstructured documents into a web of data. This would also make easier to analyze this data, because it would be already in a network form.

Semantic web is based on Uniform Resource Identifier (URI), Resource Description Framework (RDF), and Web Ontology Language (OWL). The URI is a string used to identify a name or a resource and enables interaction with representations of the resource over a network using specific protocols. RDF is a W3C standard for encoding knowledge. It is used for conceptual description or modeling of information from web resources and by computers to seek the knowledge. RDF is actually a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. The OWL is a family of knowledge representation languages for authoring ontologies.

A piece of knowledge is in RDF represented as a triple subject-predicate-object. A subject denotes the resource; the predicate denotes aspects of the resource and expresses a relationship between the subject and the object. The

resources are always named by URIs plus optional anchor IDs (URL and URN are its subsets). The triples form a multirelational network with subjects and objects represented as nodes and predicates determining types of ties – relations. There are large collections of RDF triples: *Linked Data* – Connect Distributed Data across the Web, *Freebase*, and *DBpedia*.

Different syntax formats exist and are quite varying in their complexity: N3, N-Triples, TRiG, TRiX, Turtle, RDF/XML, RDFa, and JSON-LD. The purpose of RDF is to provide an encoding and interpretation mechanism so that resources can be described in a way that a compatible software can understand it. Some formats are not human friendly but more machine friendly. See also SPARQL – an RDF query language.

Generating Random Networks

Generation of random networks (Batagelj and Brandes 2005; van der Hofstad 2011) has become important for studies of complex systems such as electrical power grid, social relations, the World Wide Web and Internet, and collaboration and citation networks of scientists. Random networks are used for modeling classes of graphs.

Paul Erdős and Alfréd Rényi proposed in Erdős and Rényi (1959) an approach to formalize the notion of a random graph. The ***Erdős-Rényi*** model, denoted by $\mathcal{G}(n,m)$, where n is the number of nodes and m is the number of edges, generates a random graph on n nodes and m edges (uniformly) randomly selected among the $n(n-1)/2$ potential edges.

Another, closely related to Erdős-Rényi model, is the ***Gilbert's*** model $\mathcal{G}(n,p)$ (Gilbert 1959), where n is the number of nodes and p is the probability that an edge is included in the random graph. In this model the $n(n-1)/2$ potential edges of a simple undirected graph $G(n,p) \in \mathcal{G}(n,p)$ are included independently with the probability p.

A model called ***small worlds*** was introduced by Watts and Strogatz (1998). This class of random graphs depends on two structural features: the clustering coefficient is high and the average distance between pairs of nodes is short. Networks such as social networks, the Internet, and gene networks all exhibit small world network characteristics.

The degree distribution of random graph from Erdős-Rényi's or Gilbert's model is sharply concentrated around its average degree. In most real-life networks, it roughly follows the powerlaw. Such networks are called ***scale-free***. Barabási and Albert (1999) described a process of ***preferential attachment*** that generates graphs with this property. The preferential attachment process creates one node at a time and each newly created node is attached to a fixed number of already existing nodes. The probability of selecting a specific node for a neighbor is proportional to its current degree.

Different classes of random graphs can be described also as ***probabilistic inductive classes*** of graphs (Kejžar et al. 2008).

Future Directions

As mentioned in the Introduction one of the basic problems in a network construction is the identification (entity resolution) problem. We expect a further development of methods and tools for solving this problem, and development and proliferation of standardized ids in different fields.

Cross-References

- Collection and Analysis of Relational Data in Organizational and Market Settings
- Ethical Issues Surrounding Data Collection in Online Social Networks
- Ethics of Social Networks and Mining
- Network Data Collected via the Web
- Network Data File Formats
- Quality of Social Network Data
- Questionnaires for Measuring Social Network Contacts

Acknowledgments This work was supported in part by Slovenian Research Agency (ARRS) – project Z7-7614 (B) and grant P1-0294, as well as by grant N1-0011 within the EUROCORES Programme EUROGIGA (project

GReGAS) of the European Science Foundation. The first author was financed in part by the European Union, European Social Fund.

References

Abdesslem FB, Parris I, Henderson T (2012) Reliable online social network data collection. Computational social networks. Springer, London

Angles R, Gutierrez C (2008) Survey of graph database models. ACM Comput Surv 40(1):1–39

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Batagelj V, Brandes U (2005) Efficient generation of large random networks. Phys Rev E 71(3):036113

Batagelj, V, Cerinšek, M: On bibliographic networks. Scientometrics 96 (2013) 3, 845–864. https://doi.org/10.1007/s11192-012-0940-1

Batagelj V, Praprotnik S (2016) An algebraic approach to temporal network analysis based on temporal quantities. Soc Netw Anal Min 6(1):1–22

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):28–37

Borgatti SP, Molina JL (2003) Ethical and strategic issues in organizational network analysis. J Appl Behav Sci 39(3):337–349

Breiger RL (2005) Ethical dilemmas in social network research: introduction to special issue. Soc Netw 27(2):88–93

Carley KM (2003) Dynamic network analysis. CASOS/CMU, Pittsburgh

Charlesworth A (2008) Understanding and managing legal issues in internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London

Erdős P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297

Eynon R, Fry J, Schroeder R (2008) The ethics of internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London

Franzosi R (2004) From words to numbers: narrative, data, and social science. Cambridge University Press, Cambridge

Gilbert EN (1959) Random graphs. Ann Math Stat 30:1141–1144

Kejžar N, Nikoloski Z, Batagelj V (2008) Probabilistic inductive classes of graphs. J Math Sociol 32(2):85–109

Lozar Manfreda K, Vehovar V, Hlebec V (2004) Collecting ego-centred network data via the web. Metodološki zvezki 1(2):295–321

Marsden PV (1990) Network data and measurement. Ann Rev Sociol 16:435–463

Marsden PV (2011) Survey methods for network data. In: Scott J, Carrington PJ (eds) The SAGE handbook of social network analysis. SAGE, London

Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation in scientific publishing. Arxiv: http://arxiv.org/abs/1304.0473. Accessed 26 Aug 2016

Mitchell JC (1969) The concept and use of social networks. In: Mitchell JC (ed) Social networks in urban situations. Manchester University Press, Manchester

Mizruchi MS, Galaskiewicz J (1993) Networks of interorganizational relations. Soc Methods Res 22(1):46–70

Popping R (2000) Computer-assisted text analysis. SAGE, London

Sampson SF (1968) A novitiate in a period of change. An experimental and case study of social relationships. PhD thesis, Cornell University

Schmidt L (2011) Using archives. A guide to effective research. Society of American Archivists, Wheaton

Shipman J, Wilson JD, Todd A (2009) Introduction to physical science, 12th edn. Cengage Learning, Boston

Ullman J, Widom J (2008) First course in database systems, 3rd edn. Prentice-Hall, Upper Saddle River

van der Hofstad R (2011) Random graphs and complex networks. http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf. Accessed 23 Aug 2016

Voorsluys W, Broberg J, Buyya R (2011) Introduction to cloud computing. In: Buyya R, Broberg J, Goscinski A (eds) Cloud computing: principles and paradigms. Wiley, New York

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

White T (2012) Hadoop: the definite guide, 3rd edn. O'Reilly Media, Sebastopol

Web References

Approximate Nearest Neighbor Library. http://www.cs.umd.edu/~mount/ANN

CAIDA (The Cooperative Association for Internet Data Analysis) Data. http://www.caida.org/data/

Centering Resonance Analysis approach proposed by Steve Corman. http://www.crawdadtech.com/

Correlates of War. http://www.correlatesofwar.org/

Data Surfing on the World Wide Web. http://it.stlawu.edu/~rlock/datasurf.html

DBpedia. http://en.wikipedia.org/wiki/DBpedia

Edinburgh Associative Thesaurus. http://www.eat.rl.ac.uk/

Enron E-mail Data. http://www.isi.edu/~adibi/Enron/Enron.htm

Ensembl. http://www.ensembl.org/index.html

Freebase. http://www.freebase.com/

GDELT. http://blog.gdeltproject.org/mapping-media-geographic-networks-the-news-co-occurrence-globe/

Genealogy Forum. http://www.genealogyforum.com/gedcom/

GO Database. http://geneontology.org/page/go-database

ICIJ – The International Consortium of Investigative Journalists (2016) Offshore leaks database. https://offshoreleaks.icij.org/pages/database
Infochimps. http://infochimps.com/
Internet Archive. http://archive.org/index.php
Internet Movie Database. http://www.imdb.com/
KDnuggets Datasets for Data Mining. http://www.kdnuggets.com/datasets/index.html
KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/
KinSource. http://kinsource.net/csac/wiki/kinsrc/KinSources/
KONECT – The Koblenz Network Collection. http://konect.uni-koblenz.de/
Linked Data – Connect Distributed Data across the Web. http://linkeddata.org/
MAG – Microsoft Academic Graph. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
Netflix. https://www.netflix.com/
Network Data Sources on Pajek's web page. http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:index
NYC Taxi and Uber Trips. http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/
Pandora. http://www.pandora.com/
Paul Hensel's International Relations Data Site. http://www.paulhensel.org/data.html
Protein Data Bank. http://www.rcsb.org/pdb/home/home.do
Public Data Sets on Amazon Web Services. http://aws.amazon.com/publicdatasets/
Repositories of Datasets. http://www.trustlet.org/wiki/Repositories_of_datasets
Revolutions: Bike sharing in 100 cities. http://blog.revolutionanalytics.com/2013/07/bike-sharing-in-100-cities.html
SNAP. http://snap.stanford.edu/data/index.html
The Internet2 Observatory Data Collections. http://www.internet2.edu/observatory/archive/data-collections.html
The Kansas Event Data System. http://web.ku.edu/keds/
TunedIT. http://tunedit.org/repo
Web Archiving Service. https://archive-it.org/