SOCIAL NETWORKS

Felix Schwagereit & Steffen Staab
ISWeb — Information Systems and Semantic Web
Universität Koblenz-Landau
http://isweb.uni-koblenz.de

DEFINITION

A social network is a social structure made of *actor*s, which are discrete individual, corporate or collective social units like persons or departments [19] that are tied by one or more specific types of *relation* or interdependency, such as financial exchange, friendship, membership in the same club, sending of messages, web links, disease transmission (epidemiology), or airline routes. The actors of a social network can have other attributes, but the focus of the social network view is on the properties of the relational systems themselves [19]. For many applications social networks are treated as graphs, with actors as nodes and ties as edges. A *group* is the finite set of actors the ties and properties of whom are to be observed and analysed. In order to define a group it is necessary to specify the network boundaries and the sampling. *Subgroup*s consist of any subset of actors and the (possible) ties between them.

The science of social networks utilizes methods from general network theory and studies real world networks as well as structurally similar subjects dealing e.g. with information networks or biological networks.

HISTORICAL BACKGROUND

The science of social network analysis comprises methods from social sciences, formal mathematical, statistical and computing methodology [19]. The first developments of scientific methods were empirically motivated and date back to the late 19th century. Jacob Moleno developed methods to facilitate the understanding of friendship patterns within small groups in the 1920's and 30's. Other pioneers in the field of social networks were Davis, who studied social circles of women in an unnamed American city and Elton Mayo, who studied social networks of factory workers. Many of the current formal concepts (e.g. density, span, connectedness) have been introduced in the 1950's and 1960's as ways to describe social structures through measures. Another important milestone was an experiment Stanley Milgram conducted in 1967. In Milgram's experiment, a sample of US individuals were asked to reach a particular target person by passing a message along a chain of acquaintances. The average length of successful chains turned out to be about five intermediaries or six steps of separation.

Early research on social networks was limited to small networks with up to a few hundred actors, which could be examined visually. With increased computational power for data acquisition and management, networks may now comprise several millions of actors.

SCIENTIFIC FUNDAMENTALS

1 Types

The simplest type of network consists of only one set of actors and one relation representing one type of ties between the actors. More complex networks can be composed of different types of actors (*multi-modal*) and different relations (*multi-relational*). Furthermore the actors and ties between them can have assigned properties, which are mostly numerical. Ties can have a direction, which makes the network a directed graph. Figure 1 shows a selection of network types [12]. Network (a) is a directed network in which each edge has a direction; (b) is an undirected network with only one type of actors; (c) is a network with several types of actors and relations; (d) shows a network with different weights for actors and ties.

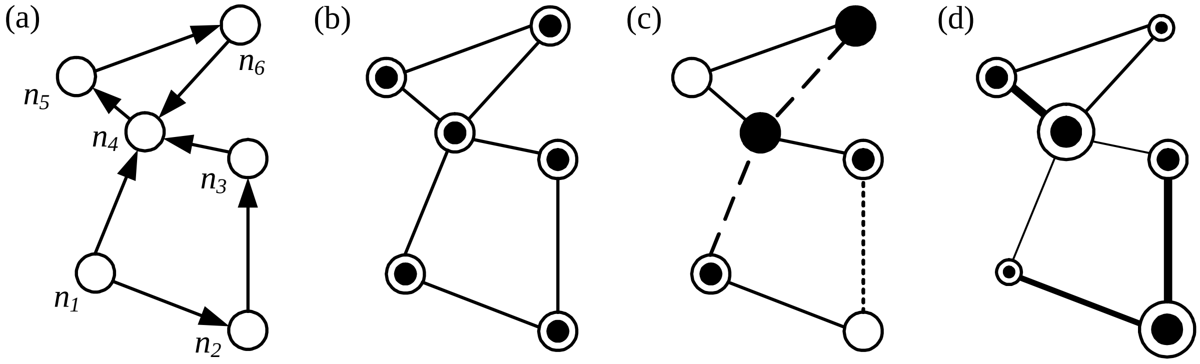

Figure 1: Types of Social Networks

Of special interest in science of social networks are *bipartite graphs* [12] which contain actors of two types and ties connecting only actors of different types. They are called *affiliation networks* because they are suitable to express the membership of people (one type of actors) in groups (the second type of actors).

2 Notation

The common notation for social networks is the *sociometric notation* [19]. Simple social networks with one relation and only one group of actors (like the one shown above) are represented as a matrix, called *sociomatrix* or *adjacency matrix*. For one relation X, we define $\mathbf{X}$ as the corresponding matrix. This matrix has g rows and g columns. The value at position x_{ij} denotes whether there exists a tie from the ith element of the social network to the jth element. An example sociomatrix for the social network (a) in Figure 1 is shown in Table 1. For more complex networks, like multi-modal and/or multi-relational social networks, tensors may be used instead of matrices [18].

	n_1	n_2	n_3	n_4	n_5	n_6
n_1	-	1	0	1	0	0
n_2	0	-	1	0	0	0
n_3	0	0	-	1	0	0
n_4	0	0	0	-	1	0
n_5	0	0	0	0	-	1
n_6	0	0	0	1	0	-

Table 1: Sociomatrix for network (a) in Figure 1

3 Measures

Measures have been developed in order to formalize local and global properties for social networks. Local and global properties of social networks describe ego-centric properties of individual actors and socio-centric properties of the network as a whole, respectively. Furthermore, we may determine subsets of actors, which we call subgroups. In the following paragraphs we provide an outline of several basic concepts.

Socio-centric Properties In order to compare different social networks in size and structure the following basic measures have been established.

Number of Actors: g

•*Number of Ties*: m

•*Mean Standarized Degree (Density)*: $z = \frac{\sum C_D(n_i)}{g(g-1)}$

•*Mean Actor-Actor distance / Characteristic Path Length*: $l = \frac{1}{\frac{1}{2}g(g+1)} \sum_{i \geq j} d(n_i, n_j)$

•*Diameter*: is the longest distance between all pairs of nodes of a given network. The distance $d(n_i, n_j)$ between a pair of nodes n_i and n_j in the network is the length of the geodesic (which is the shortest path between the two nodes).

Ego-centric Properties The identification of the "most important" or "prominent" actor was one of the primary goals of social network analysis [19]. Therefore various measures were developed to quantify "importance" of actors and subgroups for a given social network. The following measures can be calculated for simple undirected graphs of social networks.

Actor Degree Centrality is the count of the number of ties to other actors in the network. The relevance of this measure is based on the assumption that an actor, which has more connections than other actors can be considered more active and therefore important. The actor degree centrality is calculated from sociomatrix **X** as follows:

$$C_D(n_i) = \sum_j x_{ij}$$

•*Actor Closeness Centrality* is the degree to which an individual is close to all other individuals in a network (directly or indirectly). Therefore an actor is central if it can quickly (that means by relying on so few mediators as possible) interact with all other actors. The index of actor closeness centrality is:

$$C_C(n_i) = \left[\sum_{j=1, j \neq i}^{g} d(n_i, n_j) \right]^{-1}$$

where $d(n_i, n_j)$ is the length of the geodesic of actor i and actor j. To allow comparisons between different networks actor closeness can be standardized:

$$C'_C(n_i) = \frac{g-1}{\left[\sum_{j=1, j \neq i}^{g} d(n_i, n_j)\right]}$$

•*Actor Betweenness Centrality* is the degree to which an individual lies between other individuals in the network. Therefore it is based on the assumption that all other actors lying between have a certain amount of control on the interaction relying on them. So the betweenness of an actor is higher if more of the possible interactions rely on it as mediator. In order to calculate this measure we need g_{jk}, the number of geodesics linking two actors j and k; as well as $g_{jk}(n_i)$ which is the number of geodesics linking two actors that contain the actor i:

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}$$

For comparisons the measure can be normalized:

$$C'_B(n_i) = C_B(n_i)/[(g-1)(g-2)/2]$$

Subgroups In most social networks actors organize themselves in subgroups or cliques, which have their own values, sub-cultures and structures. Therefore several methods to define and recognize certain kinds of subgroups were developed [19].

A *Clique of size k* is a subgroup consisting of k many actors which are all adjacent to each other.

•An *n-Clique* is a subgroup with the property that the distance (length of the geodesic) between all actors is no greater than n and there is no actor with a distance equal or less than n outside the n-clique. An n-clique with $n = 1$ is equal to a normal clique.

•A *k-Core* is a subgroup with each actor is adjacent to at least k other actors in the subgroup.

•A *Cluster* is a subgroup consisting of actors which are similar to each other. The similarity (structural equivalence) of two actors can be defined with criteria like euclidean distance or correlation based on vectors of a sociomatrix. Similarity based clusters in undirected networks are usually created by using agglomerative or divisive hierarchical clustering methods [14]. For clustering directed networks methods like directed spectral clustering [7] can be used. In general graph theory there exist methods for partitioning graphs which can also be applied to graphs of social networks. One of these methods is the min-max cut algorithm which pursues the goal of minimizing the similarity between subgraphs while maximizing the similarity within each subgraph. Other clustering approaches are based on methods for finding densely connected subgroups by the calculation of special clustering coefficients or by comparing the number of connections within a subgroup with the number of connections to outside actors.

4 Topological Properties

Small-World Topology The small-world model [20] is a well studied distribution model of actors and ties since it has interesting properties and features. Due to the fact that networks often have a geographical component to them it is reasonable to assume that geographical proximity will play a role in deciding which actors are connected. So in a small-world network each actor is connected to actors in its near neighbourhood. Other connections between more distant actors (long-range connections) are infrequent and have a low probability. The probability for each actor of having a degree k follows a power law $p_k \sim k^{-\alpha}$ with α as constant scaling exponent. Despite the fact that long-range connections occur only sporadicly the diameter of small-world networks is exponentially smaller than their size, being bounded by a polynomial in $\log g$, where g is the number of nodes. In other words, there is always a very short path between any two nodes [8].

The discovery that real world social networks might have small-world characteristics explains the importance of this model. So it can be observed that the chain of social acquaintances required to connect one arbitrary person to another arbitrary person anywhere in the world is generally short. This concept gave rise to the famous phrase six degrees of separation after a 1967 small-world experiment by Stanley Milgram. Academic researchers continue to explore this phenomenon. A recent electronic small-world experiment [5] at Columbia University showed that about five to seven degrees of separation are sufficient for connecting any two people through e-mail. Other applications of the small-world model are investigations of iterated games, diffusion processes or epidemic processes [12].

Creation of Networks Artificially generated graphs allow comparison with real datasets and by analysing and comparing their properties they give insights into the inner structure of social networks. They also allow for the generation of (overlay) network structures on top of existing informations structures.

Several procedures are known to generate social networks from scratch. A *Poisson random graph* is the simplest way to construct a social network. This is simply done by connecting each pair of actors with the probability of p. The result of this procedure is a network with a Poisson degree distribution ($p_k = \frac{\lambda^k}{k!}e^{-\lambda}$). Since this distribution is unlike the highly skewed power-law distributions of real world networks other methods have been proposed [12]. One of the important methods is known as *preferential attachment* [1]. In this model, new nodes are added to a pre-existing network, and connected to each of the original nodes with a probability proportional to the number of connections each of the original nodes already had. I.e., new nodes are more likely to attach to hubs than peripheral nodes or in other words the "rich-get-richer". Statistically, this method will generate a power-law distributed small-world network (that is, a scale-free network).

Since there is evidence that the preferential attachment model does not show all the properties real world networks obey, like increasing of the average degree and shrinking of the diameter on growing of a network, other models have been proposed [9]. The *Community Guided Attachment*, which is based on a decomposition of actors into a nested set of subgroups, such that the difficulty of forming new links between subgroups increases with the size of the subgroups. In the *Forest Fire Model* new actors are attached to the network by burning through existing ties in epidemic fashion.

KEY APPLICATIONS

Distributed Information Management

Social routing allows to route efficiently in peer-to-peer networks without knowledge about the global network structure. This routing with local knowledge can be achieved by regarding the network as a social network and exploiting several properties of social networks like small-world characteristics. [10, 8]

Information Replication in information networks can improve scalability and reliability. By performing social network clustering on these structures prefetching of content can be improved. [15]

Information Extraction

Name disambiguation is a technique for distinguishing person names in unsupervised information frameworks (e.g. web pages), where unique identifiers can not be assumed. [2]
Ontology Extraction methods can be performed on social network structures like communities and their folksonomies. This approach is based on the assumption that individual interactions of a large number of actors might lead to global effects that could be observed as semantics. [11]

Social Recommendations

*Social networking portal*s like Xing or LinkedIn allow users to express their relationships to other users and to provide personal information. This social network can be used e.g. for finding a short path to persons in special positions by identifying the geodesic to them. [16]
Filtering, recommendations and inferred trust can be improved by taking into account the social networks all relevant actors are involved. So e.g. the trustworthiness of Bob can be inferred from a social network by Alice even if both are not directly known to each other. [6]
Viral marketing is the strategy to let satisfied customers distribute advertisements (e.g. videoclips) by recommendation or forwarding to other potential customers they know. Viral marketing campaigns are usually started by sending the advertisements to actors holding central positions in social networks in order to facilitate a rapid distribution. [13, 17]

FUTURE DIRECTIONS

For the future of the social network science many areas remain insufficiently explored [12]. Many properties of social networks have been studied in the past decades. Currently we are still lacking the whole picture which shows us what the most important properties for each application are. Especially generalized propositions (e.g. "Are more centralized organizations more efficient?") about the structure of social networks need further verification across a large number of networks [19]. Another important direction of future research is to improve our understanding of the dynamics in and the evolution of social networks [3]. In order to archive this new and more sophisticated models of social networks have to be developed. New kinds of data including more complex structures and new properties of actors or relations demand further generalization of current models. An example of these more complex structures are multiple relations which connect more than two actors.

DATA SETS

- **E**nron Email dataset (`http://www.cs.cmu.edu/ enron/`and `http://www.enronemail.com/`) contains about 600,000 Email messages belonging to 156 users. It was made public during the legal investigation concerning the Enron corporation.
- **T**he Internet Movie Data Base (IMDB) (`http://www.imdb.com/interfaces/`) is a collection of data about movies (about 400,000) and actors (about 900,000). Especially the affiliation network of the co-appearance of actors in the same movie is subject of several studies. (cf. "The Oracle of Bacon" `http://oracleofbacon.org/`)
- **D**igital Bibliography & Library Project (DBPL) collects the bibliographic information on major computer science journals and proceedings (currently about 950,000 articles). Similar to the IMDB the co-authorship can be used to generate affiliation networks. (dataset `http://dblp.uni-trier.de/xml/`)
- **S**outhern Woman Dataset, which was collected in the 1930's is published in the classical study of Davis [4], a pioneer of social network analysis. It contains the attendance at 14 social events by 18 women in an unnamed US city.

URL TO CODE

Tools and Libraries

(cf. overview on `http://www.insna.org/INSNA/soft_inf.html`):

Jung: `http://jung.sourceforge.net/`

- Pajek: `http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm`
- Ucinet: `http://www.analytictech.com/ucinet/ucinet.htm`

Conference Series

International Sunbelt Social Network Conferences: `http://www.insna.org/INSNA/sunbelt_inf.html`

Journals

Social Networks: `http://www.elsevier.com/locate/socnet`

• •CONNECTIONS: `http://www.insna.org/indexConnect.html`

•Journal of Social Structure: `http://www.cmu.edu/joss/`

CROSS REFERENCE

Scientometrics, Social Network Analysis, Network Theory, Graph Theory, Information Networks, Biological Networks

RECOMMENDED READING

Between 3 and 15 citations to important literature, e.g., in journals, conference proceedings, and websites.

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[2] Ron Bekkerman and Andrew McCallum. Disambiguating web appearances of people in a social network. In Allan Ellis and Tatsuya Hagino, editors, *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 463–470. ACM, 2005.

[3] Tim Berners-Lee, Wendy Hall, James Hendler, Nigel Shadbolt, and Daniel J. Weitzner. Creating a science of the web. *Science*, 313, August 2006.

[4] A. Davis, B. B. Gardner, and M. R. Gardner. *Deep South.* The University of Chicago Press, 1941.

[5] P.S. Dodds, R. Muhamad, and D.J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.

[6] Jennifer Golbeck and James A. Hendler. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Techn.*, 6(4):497–529, 2006.

[7] J. Huang, T. Zhu, and D. Schuurmans. Web communities identification from random walks. *Joint European Conference on Machine Learning and European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-06)*, 2006.

[8] Jon Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[9] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[10] Alexander Löser, Steffen Staab, and Christoph Tempich. Semantic social overlay networks. *IEEE JSAC — Journal on Selected Areas in Communication*, 25(1):5–14, 2007.

[11] Peter Mika. *Social Networks and the Semantic Web.* Springer, 2007.

[12] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[13] Matt Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 61–70. ACM Press, 2002.

[14] John Scott. *Social Network Analysis: A Handbook.* Sage Publications, 2000.

[15] Antonis Sidiropoulos, George Pallis, Dimitrios Katsaros, Konstantinos Stamos, Athena Vakali, and Yannis Manolopoulos. Prefetching in content distribution networks via web communities identification and outsourcing. *World Wide Web Journal*, 11(1):39–70, 2008.

[16] Steffen Staab, Pedro Domingos, Peter Mika, Jennifer Golbeck, Li Ding, Timothy W. Finin, Anupam Joshi, Andrzej Nowak, and Robin R. Vallacher. Social networks applied. *IEEE Intelligent Systems*, 20(1):80–93, 2005.

[17] Mani R. Subramani and Balaji Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Commun. ACM*, 46(12):300–307, 2003.

[18] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, New York, NY, USA, 2006. ACM.

[19] Stanley Wasserman and Katherine Faust. *Social network analysis.* Cambridge University Press, Cambridge, 1994.

[20] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.