Social Network of Co-occurrence in News Articles

Arzucan Özgür and Haluk Bingol

Department of Computer Engineering
Boğaziçi University
Bebek, İstanbul, 34342, Turkey
{ozgurarz, bingol}@boun.edu.tr

Abstract. Networks describe various complex natural systems including social systems. Recent studies have shown that these networks share some common properties. While studying complex systems, data collection phase is difficult for social networks compared to other networks such as the WWW, Internet, protein or linguistic networks. Many interesting social networks such as movie actors' collaboration, scientific collaboration and sexual contacts have been studied in the literature. It has been shown that they have small-world and power-law degree distribution properties.
In this paper, we investigate an interesting social network of co-occurrence in news articles with respect to small-world and power-law degree distribution properties. 3000 news articles selected from Reuters-21578 corpus, which consists of news articles that appeared in the Reuters newswire in 1987 are used as the data set. Results reveal that like the previously studied social networks the social network of co-occurrence in news articles also possesses the small-world and power-law degree distribution properties.

Keywords: small-world, power-law, scale-free, social-network

1 Introduction

Networks with complex topology describe various complex real world systems such as neural network of a worm (C.Elegans), power grid of the Western United States, phone-call networks, networks of linguistics, protein-folding, World Wide Web, Internet and social systems. Recent studies have shown that these networks possess some common properties such as being small-world and scale-free. Social networks describe human societies whose nodes are individual people and links represent a social interaction among these people [1]. Although obtaining data about social networks is difficult, many interesting social networks such as social network of scientific collaboration [2], movie actors' collaboration [3], sexual contacts [4], and email lists [5] have been studied in the literature. It has been shown that they share the small world concept and power law degree distribution property of scale-free networks. In this paper we investigate the small-world and scale-free properties of an interesting and previously unstudied social network of co-occurrence in news articles. In Section 2, we discuss how we have constructed this social network. In Section 3, we examine the small-world properties such as network diameter, clustering coefficient and average path length and scale-free properties such as degree distribution of the network. We conclude in Section 4.

2 Construction of the Social Network

Reuters-21578 corpus [6] consists of 21578 news articles that appeared in the Reuters newswire in 1987. This is a standard data set used extensively in research in automatic document categorization, information retrieval, machine learning and other corpus-based research. The news articles in the corpus are mostly about economics and politics.

Undirected graph of social network of co-occurrence in news articles is constructed as follows:

(i) 3000 news articles in the Reuters-21578 corpus [6] are read and person names are identified;
(ii) Nodes of the network are defined as distinct people;
(iii) A link is constructed between two people if their names appear in the same news article.

The social network constructed consists of 459 nodes and 1422 edges. To analyze and visualize the network Pajek Network Analysis and Visualization Program [7] is used. The graph of the constructed network is presented in Figure 1. It can be seen from the graph that there are many vertices that have 1 or 2 connections such as Gerhard de Kock, who was the Governor of the Central Bank of South Africa in 1987, Ferdinand Lacina, who was the Finance Minister of Austria in 1987, and Subroto, who was the Mining and Energy Minister of Indonesia in 1987. We can also observe that there are vertices which are very highly connected such as Ronald Reagan, the President of the USA in 1987 and James Baker, the Treasury Secretary of the USA in 1987.

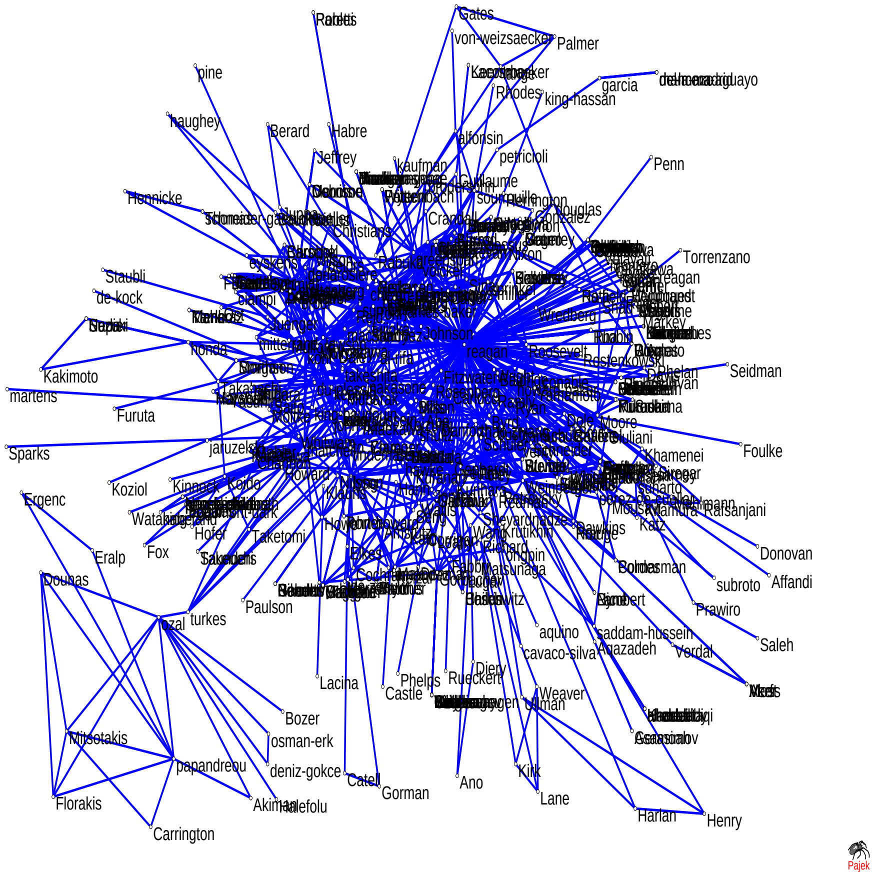

Fig. 1. Visualization of the social network of co-occurrence in news articles

3 Characteristics of the Social Network of Co-occurrence in News Articles

In this section, diameter, clustering coefficient, average path length and degree distribution properties of the network will be discussed. An early model for complex graphs is the Random-Graph model, which is due to the work of Erdos and Renyi [8]. In Random Graphs, a large number of nodes are randomly connected and the properties of the graph as the number of nodes goes to infinity are studied. Properties observed in real world complex networks are compared with respect to that of random graphs since they constitute a measure.

3.1 Small-World Property

Small-world concept describes the fact that despite their large size, most real world networks have a relatively short distance between any two nodes. *Distance* between two nodes is defined as the number of edges along the shortest path between them. *Diameter* of a network is the longest of the shortest distances between all pairs of nodes in the network. The property of small-world networks is that they have small diameter and small average path length between vertices. The most popular example of small world networks is the "six-degree-of-separation" concept uncovered by the social psychologist Stanley Milgram in 1967 [9]. Stanley Milgram concluded that there is a path of acquaintances with a typical length of six between most pairs of people in the United States [9]. Other small-world network examples are actor-movie network and chemicals in a cell network [9]. Actors in Hollywood are on average three co-stars apart from each other, when generalized to the actors all over the world this distance increases to six. Likewise, chemicals in a cell are separated typically by three reactions.

In order to observe whether the small-world concept is valid for our social network of co-occurrence in news articles, we have calculated its diameter and average path length. The diameter of the network is calculated to be 8 due to the path from Oil, Mines and Parastatal Industry Minister of Mexico in 1987 Alfredo Del Mazo to the Turkish Foreign Ministry spokesman Yalim Eralp in 1987.

The average path length of the network is calculated to be 2,98. The diameter and average path length of the network are relatively small compared to the size of the network. Thus, we can conclude that this is a small world network.

3.2 Clustering Coefficient

A property of social networks is that cliques exist, where there are circles of individuals that all know each other. This property is quantified by the clustering coefficient. *Clustering coefficient of a node i* is defined as [9]:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{1}$$

Here, k_i is the degree of node i and E_i is the number of links between the k_i neighbors of node i. *Clustering coefficient of the whole network* is the average of all C_i's. We have calculated the clustering coefficient of the network as 0,02. The clustering coefficient of a random network with the same size and average degree generated according to the Erdos-Renyi Model [8] is calculated to be 0.003, which is an order of magnitute smaller than our social network. This is another indicator that our network has small-world property.

3.3 Degree Distribution

Degree of a node is the number of edges the node has. The spread of the degrees of nodes in a network is characterized by a distribution function $P(k)$. $P(k)$ is the probability that a randomly selected node has k edges. In random graphs, $P(k)$ follows a Poisson distribution which has a peak at the average degree of the network $\langle k \rangle$. Therefore majority of the nodes have the same degree around $\langle k \rangle$ and extremely few nodes have very small or very large degrees. In the recent years, studies have shown that most real world networks such as World Wide Web [10], Internet [11], metabolic networks [12] and social networks such as movie actors [3], coauthor networks [2], sexual contacts network [4] follow power law degree distribution characterized as:

$$P(k) \sim k^{-\gamma} \tag{2}$$

Here γ is called the scaling factor and such networks are called scale-free [9]. Power law distribution implies that nodes with few links are numerous, while very few nodes have very large number of links. In this study, we observed the degree distribution of the social network of co-occurrence in news articles. In Table 1, the degree distribution is given. Since in social networks persons are important, a representative person for each frequency is also given. Average degree of the network is calculated to be $\langle k \rangle = 6.02$.It is observed that most of the nodes have degrees less that 10. Very few nodes have large degree values. The most connected node is the node representing Ronald Reagan, with degree 209. This node acts as the hub of the network since 45% of the nodes are connected directly to it.

Table 1. Frequency of the degrees of nodes and a representative node for each degree

Degree	Frequency	Representative (Position)
1	74	Von-Weizsaecker (President of West Germany)
2	80	Haughey (Prime Minister of Ireland)
3	72	Ongpin (Finance Minister of the Philippines)
4	28	Alfonsin (President of Argentina)
5	52	Brodersohn (Finance Secretary of Argentina)
6	29	Wilson (Finance Minister of Canada)
7	35	Conable (World Bank President)
8	21	de-Larosiere (Central Bank of France Governor)
9	13	Camdessus (International Monetary Fund Managing Director)
10	10	Leigh-Pemberton (Central Bank of England Governor)
11	10	Sprinkel (Chairman of Council of Economic Advisers)
12	5	Reid (Reid-Ashman Inc Company Founder)
13	2	Schlesinger (Vice-President of Central Bank of West Germany)
14	2	Carter (Former President of USA)
15	4	Sumita (Central Bank of Japan Governor)
16	4	Lawson (Finance Minister of UK)
18	1	Gorbachev (President of USSR)
19	1	Balladur (Minister of State for Economy, Finance and Privatization of France)
20	1	Shultz (U.S. Secretary of State)
25	1	Kohl (Chancellor of West Germany)
26	1	Poehl (President of Central Bank of West Germany)
28	1	Miyazawa (Finance Minister of Japan)
29	1	Gephardt (Representative, Missouri Democrat, USA)
30	1	Yeutter (Representative for Trade Negotiations of USA)
31	2	Greenspan (Chairman of Federal Reserve Board of USA)
32	1	Howard-Baker (White House Chief of Staff)
35	1	Lyng (Agriculture Secretary of USA)
37	1	Volcker (Federal Reserve Board Chairman of USA)
42	1	Stoltenberg (Finance Minister of West Germany)
43	1	Thatcher (Prime Minister of UK)
58	1	Nakasone (Prime Minister of Japan)
77	1	James-Baker (Treasury Secretary of USA)
209	1	Reagan (President of USA)

Graph of the degree distribution of the network is drawn in the logarithmic scale as in Figure 2. We can see that this degree distribution does not follow Poisson distribution, but follows power-law distribution. Scaling factor γ is calculated to be $1,7$.

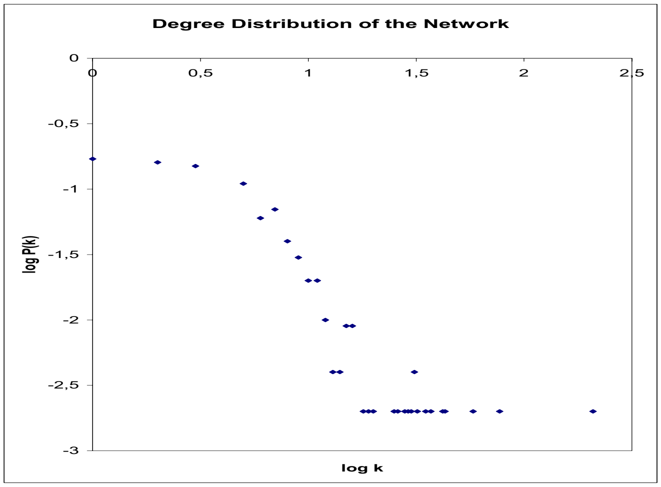

Fig. 2. Degree distribution of the social network of co-occurrence in news articles

3.4 Summary of General Characteristics of the Network Together With Other Real World Social Networks

In Table 2, summary of the properties of the social network of co-occurrence in news articles is given together with some other previously studied social networks. $Size$, is the number of nodes in the network, $\langle k \rangle$ is average degree of the nodes of the network, $\langle l \rangle$ is average path length between a pair of nodes in the network, C is the clustering coefficient of the network, $C(rand)$ is the clustering coefficient of a random network with same size and same number of edges, and γ is the scaling factor of the network.

Common properties of all the networks listed in Table 2 are that they have relatively small average path length and diameter compared to their sizes; they follow the power-law degree distribution and have relatively high clustering coefficient compared to the random networks with same sizes and same average degrees.

Table 2. General properties of some social networks

Network	Size	$\langle k\rangle$	$\langle l\rangle$	C	$C(rand)$	γ
Coauthors, SPIRES [9]	56 627	173,0	4,0	0,726	0,003	1,2
Social network of co-occurrence in news articles	459	6,02	2,98	0,02	0,003	1,7
Coauthors, neuroscience [9]	209 293	11,5	6,0	0,76	0,000055	2,1
Movie Actors [3]	225 226	61,0	3,65	0,79	0,00027	2,3
Coauthors, Math. [9]	70 975	3,9	9,5	0,59	0,000054	2,5
Sexual Contacts [4]	2 810	-,-	-,-	-,-	-,-	3,4

4 Conclusion

In this paper we investigated the small-world and scale-free properties of an interesting social network not studied previously. 3000 news stories mostly related to economics and politics that appeared in the Reuters newswire in 1987 are read and people names are extracted. The Reuters-21578 corpus from where the news stories are read is a standard data set used widely in information retrieval, machine learning and document categorization. A social network is constructed from this data. Nodes of the network are distinct people. There is a link between two people if they have appeared in the same news story. The resulting network is an undirected network composed of 459 nodes and 1422 links.

We have observed that this network has a relatively small diameter and average path length compared to its size. It has also relatively large clustering coefficient compared to the clustering coefficient of random network of the same size and average degree. The degree distribution of the network is also studied and it has been shown that it follows a power law distribution. There are numerous nodes with small degrees and very few nodes with high degrees. The most connected node is the node representing Ronald Reagan, the President of USA in 1987. 45% of the nodes are directly connected to that node and it plays the role of a hub in this network. It is concluded that like the previously studied social networks such as movie-actor collaboration network, co-authorship network and the sexual contact network the co-appearance in news article social network possesses the small world property and power low degree distribution property of the scale-free networks. As feature work, we will extend our network to cover more articles from the Reuters-21578 data set and observe the properties of this extended network. We expect to use our approach as a function of search engines to find related people. This needs further study.

Acknowledgments

This work has been partially supported by Boğaziçi University Research Fund under the grant number 33HA101.

References

1. Barabsi, A. L., Deszo Z., Ravasz E., Yook S. H., Oltvai Z.: Scale-free and hierarchical structures in complex networks. To appear in Sitges Proceedings on Complex Networks. (2004)
2. Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert A., Vicsek, T.: Evaluation of the social network of scientific collaborations. Physica A (2002) 590–614
3. Watts, J. D., Strogatz, S. H.: Collective dynamics of small-world networks. Nature **393** (1998) 440–442
4. Liljeros, F., Edling, C. R., Amaral, L. A. N. , Stanley, H. E., Aberg, Y.: The web of human sexual contacts. Nature **411** (2001) 907–908
5. Kirlidog, M., Bingol, H.: The shaping of an electronic list by its active members. 5th International IT in Regional Areas Conference (2003), Central Queensland, Australia, 40–48
6. Lewis, D. D.: Reuters-21578 corpus. Available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.
7. Pajek. Package for Large Network Analysis. Available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/
8. Erdos, P., Renyi, A.:On Random Graphs I. Publ. Math. Debrecen **6** (1959) 290–297
9. Albert, R., Barabasi, A. L.: Statistical mechanics of complex networks. In Reviews of Modern Physics **73** (January 2002) 47–97
10. Adamic, A. L., Huberman, B. A.: Growth dynamics of the World Wide Web. Nature **401** (1999) 131
11. Yook, S., Jeong, H., Barabasi, A. L.: Modeling the Internet's large-scale topology. PNAS **99** (2002) 13382–13386
12. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabasi, A. L.: The large-scale organization of metabolic networks. Nature **407** (2000) 651

This article was processed using the LaTeX macro package with LLNCS style