Co-occurrence Network of Reuters News

Arzucan Özgür,[1] Burak Cetin,[2] and Haluk Bingol[2]

[1]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA*
[2]*Complex Systems Research Laboratory, Department of Computer Engineering, Bogazici University Istanbul, Turkey*

Networks describe various complex natural systems including social systems. We investigate the social network of co-occurrence in Reuters-21578 corpus, which consists of news articles that appeared in the Reuters newswire in 1987. People are represented as vertices and two persons are connected if they co-occur in the same article. The network has small-world features with power-law degree distribution. The network is disconnected and the component size distribution has power law characteristics. Community detection on a degree-reduced network provides meaningful communities. An edge-reduced network, which contains only the strong ties has a star topology.

"Importance" of persons are investigated. The network is the situation in 1987. After 20 years, a better judgment on the importance of the people can be done. A number of ranking algorithms, including Citation count, PageRank, are used to assign ranks to vertices. The ranks given by the algorithms are compared against how well a person is represented in Wikipedia. We find up to medium level Spearman's rank correlations. A noteworthy finding is that PageRank consistently performed worse than the other algorithms. We analyze this further and find reasons.

I. INTRODUCTION

Networks describe various complex real world systems such as neural network of a worm (Caenorhabditis elegans), power grid of the Western United States, phone call networks, networks of linguistics, protein-folding, the World Wide Web and social systems [1]. Social networks describe human societies whose nodes are individual people and links represent a social interaction among these people. 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], email lists [5], and terrorist network of September 11, 2001 [6] have been studied in the literature. Recent studies have shown that these networks possess some common properties such as being small-world and scale-free.

Reuters-21578 corpus is a standard data set used extensively in research in automatic document categorization, information retrieval, machine learning and other corpus-based research [7]. It consists of $21,578$ news articles, mostly about economics, that appeared in the Reuters newswire in 1987. In a previous study, $3,000$ of these news articles are randomly selected and person names are manually identified by reading [8]. An undirected, un-weighted network where nodes are people and there is a link if two people have appeared in the same news article is constructed. Then, the scale-free and small world properties of the network is studied.

In this study we extend the network to cover all the $21,578$ news articles. We investigate component structure, community structure, co-occurence in time, and the "importance" of people.

II. CONSTRUCTION OF THE SOCIAL NETWORK

The most challenging part of the study was constructing the network [9], since it required a lot of manual effort. An undirected and weighted graph of social network of co-occurrence in news articles is constructed as follows:

1. All the news articles in the Reuters-21578 corpus are processed and person names are identified. Person names are identified manually in $10,000$ of the articles. To identify the names in the remaining articles, we used the name finder tool of Bikel et al. [10]. The output of the tool is post-processed manually to eliminate the errors.

2. Nodes of the network are defined as distinct people.

3. A link is constructed between two people if their names appear in the same news article. Each edge is associated with a weight corresponding to the number of times the two people appeared in the news stories together.

Note that in this representation there is an edge in between two persons if they co-occur in one article independent of the number of co-occurrence. Therefore, the graph represents connectivity but it could be misleading in terms of "closeness" of the people. Another observation about the representation is that the degree of a vertex could be misleading. Suppose there are 5 people in an article which contribute 4 degrees to each corresponding vertex in the graph. Probably this situation is quite different than 5 persons co-occurring pairwise in different articles although the graph representation is the same for both cases. The vertex is not weighted. Either a person is in one single article or she is in many articles, the representation is the same. The social network constructed consists of $5,249$ nodes and $7,528$ edges.

III.   CONNECTED COMPONENTS

The social network of co-occurrence in news articles turns out to be an unconnected graph. The graph consists of $2,105$ components. There is a giant component of $1,682$ vertices and $5,008$ edges. The rest of the components are small with sizes in the range of $[1, 26]$. Table I displays component size, number of occurrences of that component size in the network, number and percentage of people who appear in a component of the specified size.

TABLE I: Component size frequency table.

Size s	Frequency $p(s)$	% Frequency	# of Persons	% of Persons
1	1363	64.75	1363	25.97
2	442	20.42	884	16.84
3	138	6.56	414	7.89
4	71	3.37	284	5.41
5	35	1.66	175	3.33
6	19	0.90	114	2.17
7	11	0.52	77	1.47
8	8	0.38	64	1.22
9	5	0.24	45	0.86
10	5	0.24	50	0.95
11	3	0.14	33	0.63
12	2	0.10	24	0.46
14	1	0.05	14	0.27
26	1	0.05	26	0.50
1682	1	0.05	1682	32.04

64.75% of the components are of size 1. This means that $1,363$ people in the network are not connected to anyone, i.e., they appear just by themselves in the news. These people are generally US senators, and managers or chairs of certain corporations. In the news articles they appeared by expressing their ideas about a phenomenon or by being elected for a certain position. Examples of such patterns are "person A told" and "person A was elected to". There are 442 components of size 2. In other words, there are 442 pairs of people who are connected only to each other. Patterns of news that form components of size 2 are as "person A said and person B said" and "person A was chosen to replace person B". All of the components of size less than and equal to 26 possess similar characteristics.

We investigate the frequency $p(s)$ of occurrence of components of size s. We observed that $p(s)$ is related to s by a power law given by $p(s) \propto s^{-\gamma}$, where $\gamma = 2.8$.

In the rest of the study, the results are computed based on the giant component.

IV.   COMPLEX NETWORK ANALYSIS

The diameter, that is, the largest distance between two nodes, of the network is calculated to be 13. This is between *Jack Sandner* (Chairman of Chicago Mercantile Exchange) and *Richard Simon* (analyst at Goldman Sachs and Co). Interestingly, these people are both from USA. So, although they are geographically close to each other, they are the most distant people in the network of news. The diameter of a random graph with the same number of nodes and edges is 15.

The average path length and clustering coefficient are given in Table II. They are calculated by Pajek [11]. The diameter and average path length of the network are relatively small compared to the size of the network. The clustering coefficient is ~ 474 times greater than the clustering coefficient of a random network with the same number of nodes and edges. Thus, we can conclude that this is very different from a random graph.

The degree distribution of the network is power-law with exponent $\gamma = 2.4$. 90.66% of the nodes have degrees less that 10. The node with the maximum degree is the node representing Ronald Reagan which acts as the hub of the network since 23.50% of the nodes are connected directly to it. Other highly connected nodes in decreasing order are given in Table III.

TABLE II: Comparison of co-occurrence network with random network with the same number of vertices and edges.

Metric	Co-occurrence Network	Random Network
Diameter	13	15
Average path length	4.39	6.73
Clustering coefficient	0.6906	0.0014

TABLE III: Persons with the highest degree.

Person	Degree	Note
Ronald Reagan	395	President, USA
James Baker	158	Treasury Secretary, USA
Yasuhiro Nakasone	99	Prime Minister, Japan
Margaret Thatcher	52	Prime Minister, UK
George Bush	52	Vice President, USA
Gerhard Stoltenberg	50	Finance Minister, West Germany
William Casey	32	Head of CIA, USA
Mikhail Gorbachev	28	President, USSR

V.   COMMUNITIES

Nodes in many real networks appear to group in sub-graphs where the density of connections among the nodes in the sub-graph is larger than that of the connections with the rest of the nodes in the network. These sub-graphs are called *communities*. Since the community structure of the network is unknown, applying a community detection to this network would not be very useful since we cannot process the communities manually. We decided to reduce the graph size by removing the vertices with degree less than 10 and obtained manageable graph of 49 people. We use the Girvan-Newman betweenness clustering algorithm on the reduced network to detect the communities [12]. We identified four communities. When we manually investigate the properties of the people in the communities (based on their job positions in 1987), we observe meaningful communities.

In the first community of 21 persons, we see US President Reagan, the prime minister of Japan, US senators, White House spokesmen, and news reporters. We also see that people involved in US national security are in this community. Current and former heads of CIA, Senate Armed Services Committee Chairman, current and former national security advisors, defense secretary of USA and two news reporters are clustered in this community. Interestingly, Marine Lt. Col. Oliver North, one of the main actors of "Iran-Contra Affair" [13] political scandal, is also in this community.

There are 8 persons in the second community where we see US democratic senators, former president of USA Franklin Roosevelt, Secretary of the Commerce Department, and American Enterprise Institute analyst. It is interesting that we have a baseball manager in this community.

In the third community, there are 9 persons many of which are related to Federal Reserve such as the former and current Federal Reserve board chairmen and vice chairman, staff director of monetary fund, New York Federal Reserve Bank President, and Federal Deposit Insurance Corp. Chairman. Former president of USA Jimmy Carter is also in this community.

In the fourth community we see the Vice President of USA (later becomes president) George Bush, US democratic senators, commerce secretary, and secretary of state. The Canadian President and foreign secretary are also in this community.

VI.   STRONG TIES

We want to investigate the people connected strongly. In order to manually check, we need to reduce the network size, again. This time the edges with weight less than 10 are removed. The rationale is that if two people appear in only a few articles together they have a weaker relationship compared to people who often appear together therefore have high edge weight. So, we are left with edges that have high weight and associate these edges with representing strong relationships. Fig. 1 shows the cluster of the network. It is interesting to note that the topology is similar to a star connection, where Ronald Reagan is in the center. The highly weighted connections are between Ronald Reagan and other people in the White House such as Paul Volcker and James Baker; or presidents and politicians from other
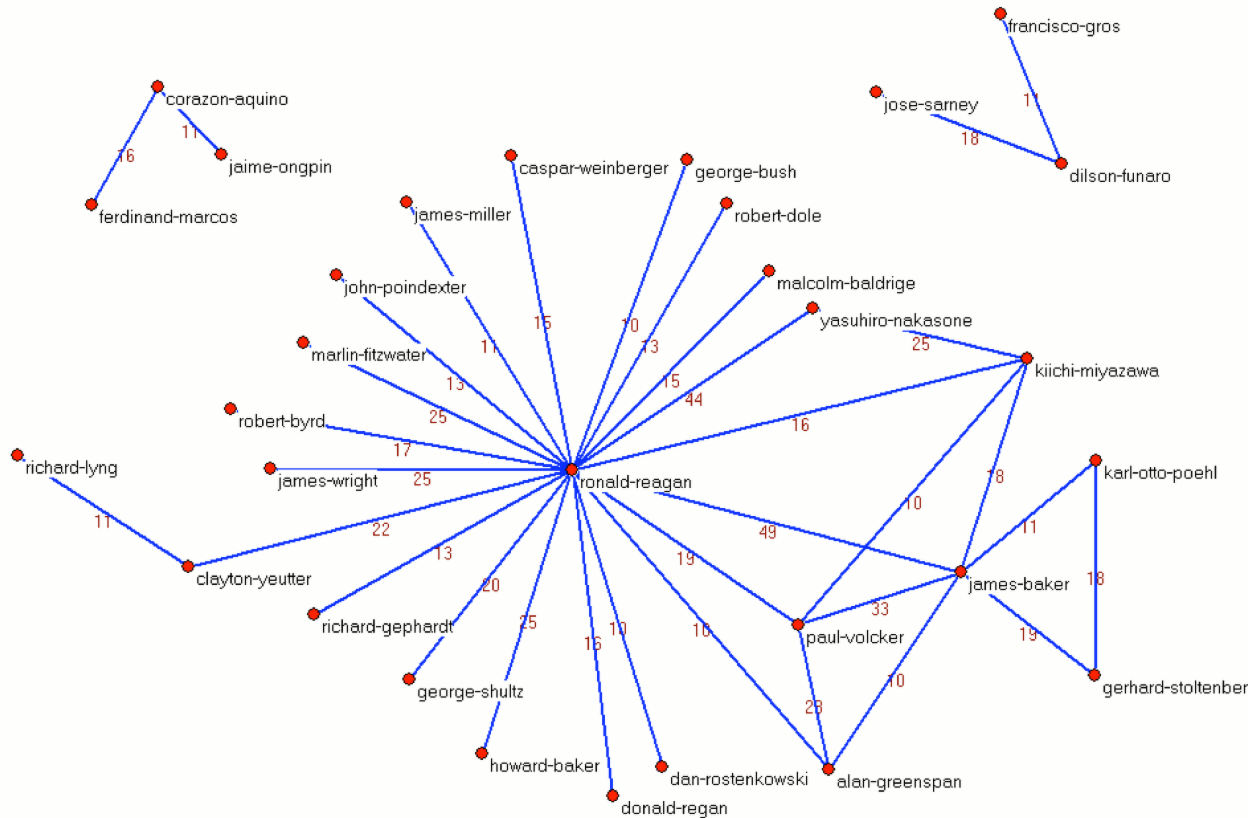
FIG. 1: Strong ties

countries such as Yasuhiro Nakasone (President of Japan).

There are two 4-cliques: {Ronald Reagan, Paul Volcker, James Baker, Alan Greenspan} and {Ronald Reagan, Paul Volcker, James Baker, Kiichi Miyazawa}. There are two triangles: {Ronald Regan, Kiichi Miyazawa, Yasuhiro Nakasone} and {James Baker, Karl Otto Poehl, Gerhard Stoltenber}.

We also note two triads that are unconnected from the star sub-network. The triad on the top left are people from Philippines {President, former president, Finance Minister} and the triad on the top right are people from Brazil {President, Finance Minister, Chairman of a bank}.

VII.   TIME SERIES OF CO-OCCURRENCE

We used the chronological order property of the news articles to compute and visualize the patterns in the time-series of occurrences of people in the news. 21,578 articles are chronologically sorted and binned into 43 bins of 500 articles each based on their date.

We graph the results. The x-axis, representing the time, is the bin number and the y-axis is the frequency of occurrence of the person in the news at the corresponding bin. The time series in Fig. 2 (a), shows the occurrence of Ronald Reagan (President of USA), Yasuhiro Nakasone (President of Japan), and Kiichi Miyazawa (Finance Minister of Japan). We note the frequency of occurrence of Ronald Reagan never drops to 0. In general, what we observe is that, for popular people (presidents, politicians, and etc.) there is a continuous occurrence throughout the period. Some other people (maybe less popular in this sense) make a peak at some point, but later disappear and never re-occur again. There is an interesting peak at bin 35. Due to space limitation the figure is not provided but when we look at some other people, James Baker and Paul Volcker also make a peak at that time point. On the other hand, people from other countries do not have peaks at that point. These patterns might show us that there is an important event going on between Japan and USA at that time point.

Another pattern that reveals an event between two countries is shown in Fig. 2 (b). Turgut Ozal is the president of Turkey and Andreas Papandreou is the president of Greece. At the time period that both presidents make peaks there was a crisis between the two countries.

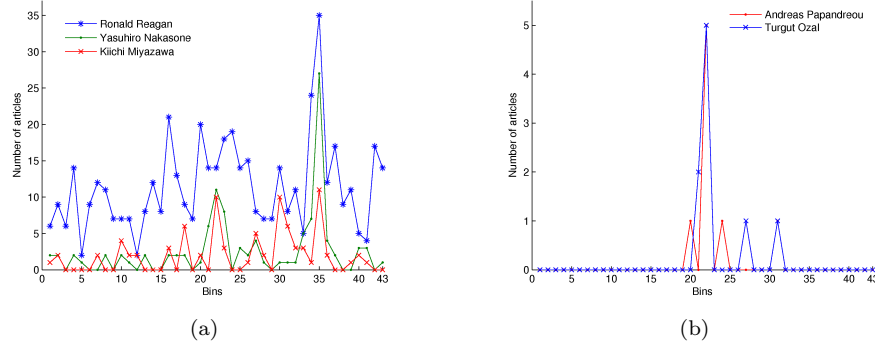(a)   (b)

FIG. 2: The frequency of occurrence changes in time. (a) The frequency of occurrence graphs of Ronald Reagan, Yasuhiro Nakasone and Kiichi Miyazawa present persistent continuous activation. (b) The frequency of occurrence graphs of Turgut Ozal and Andreas Papandreou present partial activation based on some events.

VIII.   FINDING IMPORTANT PERSONS IN THE NETWORK

The co-occurrence network contains information about the importance of people. We investigate this by applying various vertex ranking measures on our network for a comparative study. We expect to obtain the importance of persons as viewed from a Reuters perspective.

A.   Analysis Setup

In our study we will link the importance of a person in 1987 to importance today. We will see how well a person in the collection is represented in today's English Wikipedia [14] and compare that with rankings produced by various well-known centrality measures. These are the algorithms:

1. *Article count*, denoted by *acount*, is the number of articles a person appears in.

2. *Degree* is the number of people a person got associated with in the collection, i.e. the link count on the node.

3. *PageRank*, denoted by *pr*, is the PageRank of a node as described in Page et. al. [16], using $d = 0.5$. For application we have converted the undirected network to directed by replacing each edge with arcs in both directions.

4. *Closeness*, denoted by *cls*, calculated using the undirected unweighted network. *Closeness* calculates the shortest distance from each person in the network to each other person based on the links between all members of the network [15]. Here central nodes are the ones which are closest to all other nodes.

5. *Betweennes*, denoted by *btw*, calculated using the undirected unweighted network. *Betweenness* examines the extent to which a node is situated between others in a network [15]. It is a measure of how much damage to connectivity there would be if a given node is removed from the network.

For assessing the validity of our results we have used a crawler to look up if a given person has an English Wikipedia page. We have interpreted this as an indication that a given person is important today in a general global sense. This would have an English speaking world bias and may not necessarily be a truly objective measure. However Reuters also being an English source and English being the closest there is to a truly global language, this measure should function at least to a reasonable extent. Our basic assertion here is that if a person was important back in 1987 when the Reuters articles were being published, then s/he would still be important today. The 20 years passed since then can make a "time's judgment" on who were truly important at the time. It is possible however other people

in those articles unimportant or unforeseeable at the time will have gained importance. Similarly some who were not very important from a Reuters reporting perspective can actually be important individuals for different reasons. Combined, these would mean that the assessment power of the algorithms would be limited in discovering all those who are important, however this analysis should be reasonably good enough to penalize "false positives" which the algorithms would mark as important but were really not as such.

Using the crawler results we have constructed two functions:

1. *"has a page"*: $H(i) = 1$ if there is any Wikipedia page for a given person i. $H(i) = 0$ otherwise.

2. *"is a wikified page"*: $W(i) = 1$ if there is a Wikipedia page with the "Wikified" format for person i which means the page is updated according to Wikipedia biography standards e.g. has a v-card biography box giving basic details. $W(i) = 0$, otherwise.

Of the 5,249 persons in the network we find that 1,440 have a Wikipedia page and 383 have a "Wikified" page. In the rest of this section we will use these functions as apriori information on the importance of nodes and perform a comparative study of the algorithms. Table IV shows the top 20 people when ranked according to *account* values. Having a glance at this table can serve as a basic reality check for the utility of our defined functions. For example we see that many people we could expect to have high importance have $W(i) = 1$; President of USA, Prime Minister of Japan, Secretary of State of USA. Table V shows a similar table for *cls*.

TABLE IV: Top-20 persons in article count.

person	account	$H(i)$	$W(i)$	notes
ronald-reagan	493	1	1	President, USA
james-baker	212	1	1	Treasury Secretary, USA
yasuhiro-nakasone	112	1	1	Prime Minister, Japan
paul-volcker	109	1	1	Chairman of Federal Reserve Board, USA
kiichi-miyazawa	86	1	1	Finance Minister, Japan
clayton-yeutter	85	1	0	Trade Representative, USA
nigel-lawson	66	1	1	Chancellor of the Exchequer, UK
dilson-funaro	58	0	0	Finance Minister, Brazil
richard-lyng	57	1	0	Agriculture Secretary, USA
gerhard-stoltenberg	55	1	0	Finance Minister, W. Germany
george-shultz	50	1	1	Secretary of State, USA
margaret-thatcher	50	1	1	Prime Minister, UK
edouard-balladur	48	1	1	Finance Minister, France
james-wright	47	1	0	White House Speaker, Texas Democrat, USA
satoshi-sumita	44	0	0	Bank of Japan Governor, Japan
malcolm-baldrige	42	1	0	Commerce Secretary, USA
marlin-fitzwater	40	1	0	White House Spokesman, USA
alan-greenspan	39	1	1	Chairman of Federal Reserve Board, USA
jaime-ongpin	36	0	0	Finance Secretary, Philippines
jose-sarney	36	1	1	President, Brazil

B. Analysis of Algorithm Performances

In Fig. 3(a) we show the moving averages (using a window size of 500 nodes) of the algorithms for the value of $H(i)$ when nodes are sorted by their ranks in the corresponding algorithms in decreasing order. In producing this graph we randomize the positions of the nodes with equal ranks and take the average value of $H(i)$ corresponding to the position. This produces a near straight line for groups of nodes with equal ranks (e.g. the tails of the plots). The corresponding graph for $W(i)$ is Fig. 3(b).

Looking at these graphs we would like an algorithm ideally to put all the positive nodes (i.e. $H(i) = 1$ or $W(i) = 1$) at the beginning (high ranked positions) and the rest to the tail achieving a separation. Given this we see that PageRank and betweenness perform badly as these usually stay below the other algorithms for higher ranked positions. At the

TABLE V: Top-20 persons in closeness.

person	cls	H(i)	W(i)	notes
ronald-reagan	0.136	1	1	President, USA
james-baker	0.120	1	1	Treasury Secretary, USA
paul-volcker	0.114	1	1	Chairman of Federal Reserve Board, USA
george-shultz	0.112	1	1	Secretary of State, USA
yasuhiro-nakasone	0.112	1	1	Prime Minister, Japan
kiichi-miyazawa	0.110	1	1	Finance Minister, Japan
gerhard-stoltenberg	0.109	1	0	Finance Minister, W.Germany
nigel-lawson	0.109	1	1	Chancellor of the Exchequer, UK
clayton-yeutter	0.109	1	0	Trade Representative, USA
helmut-kohl	0.108	1	1	Chancellor, W.Germany
edouard-balladur	0.107	1	1	Finance Minister, France
malcolm-baldrige	0.107	1	0	Commerce secretary, USA
marlin-fitzwater	0.107	1	0	White House Spokesman, USA
howard-baker	0.106	1	1	White House Chief of Staff, USA
margaret-thatcher	0.106	1	1	Prime Minister, UK
david-mulford	0.106	1	0	Assistant Treasury Secretary, USA
barber-conable	0.106	1	0	World Bank President
brian-mulroney	0.106	1	1	Prime Minister, Canada
richard-lyng	0.106	1	0	Agriculture Secretary, USA
alan-greenspan	0.106	1	1	Chairman of Federal Reserve Board, USA

tail we see betweenness staying above the others which is bad for the same reason that it can not separate positive nodes from the rest. Closeness appears to uncover important people earlier and more intensively than others.

An interesting observation we make on both plots in Fig. 3 is the bump PageRank creates close to the middle of the graph with unexpected highs for relatively low PageRanks. Investigation reveals that right of this bump are persons who appear in a Reuters article alone with no links thus have very low PageRanks. In contrast, to the left of this bump are some overrated persons who also appear in just one article but appear with more than one person resulting in a fully connected sub-graph feeding back each other's ranks. This directly relates to the original "rank sink" problem in which unreasonably high ranks accumulate in loops without any other outgoing links but have incoming links [16]. This problem was partially remedied by introducing the random jumps for the random surfer and removing the nodes with no outgoing links for the rank calculation. We can see that the former remedy would not be very effective on this network because the conversion from undirected to directed network creates so many loops in the network. Also the latter solution would have no effect in this ranking as those nodes without any links would be place at the tail either way.

The cumulative plots in Fig. 4 plot the total number of important people ranked in the top-t people as ranked by an algorithm. Similar to the moving average plots we randomize and take averages for nodes with equal ranks. We see in these plots that all the algorithms do comparably in the top-500 persons, however later on PageRank and *betweenness* perform badly whereas closeness appears to outperform the others which is consistent with our initial observation above. For example in Fig. 4(b) we see that if we were to query for the first 1400 persons according to the algorithms, in the results PageRank would have returned about 40% of all the people with a Wikified page whereas *closeness* would have returned more at about 60%.

Since the data characteristics of the functions $H(i)$ and $W(i)$ and the rankings produced by the algorithms are different it is not appropriate to do linear correlation. Spearman correlation only requires the orderings produced by two functions to match. Spearman correlation results can be seen on Table VI. Note that we have also normalized the results to take into account that maximum (self) correlations for $H(i)$ and $W(i)$ are rather low at 0.597 and 0.203 respectively. We observe that $H(i)$ produces weak (normalized) correlations and $W(i)$ produces medium (normalized) correlations at their highest. The comparative performance we see in these results appear consistent with our observations on the cumulative graphs of Fig. 4.

The low to medium correlations are not unexpected. As we have argued earlier the algorithms are not able to discover all the important persons since that information is not necessarily there, however we should be able to see the differences in their effectiveness in disclosing the evidence that is encoded in the link structures, and this we can observe in the results.
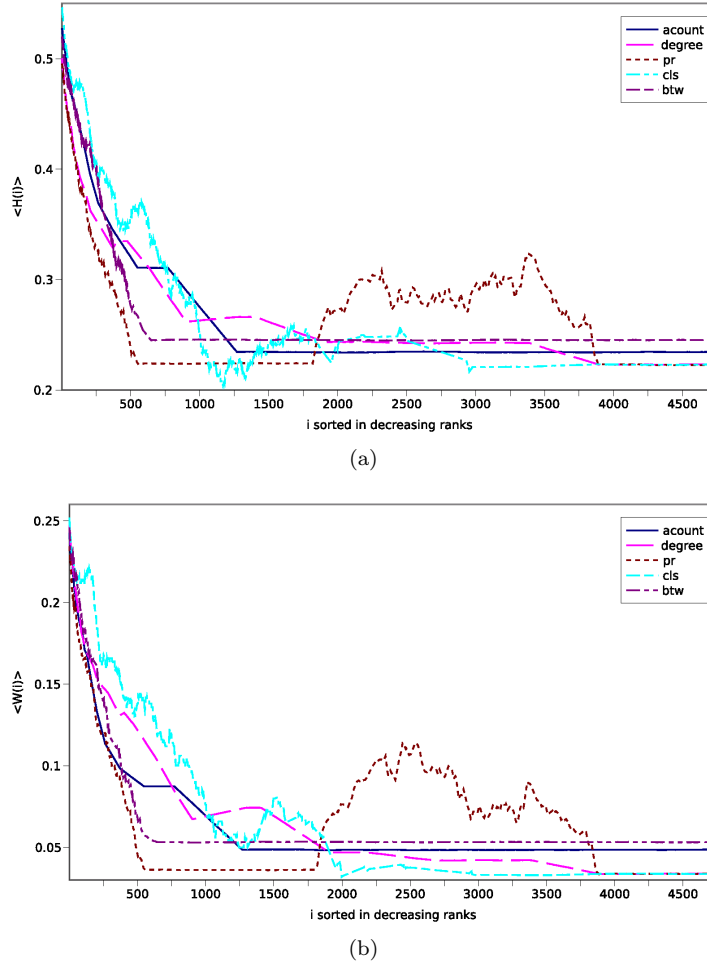
(a)

(b)

FIG. 3: (Color online) Moving averages of Wikipedia functions with window size 500 for the algorithms (a) $H(i)$, (b) $W(i)$.

TABLE VI: Spearman correlations.

algorithm	$H(i)$	$H(i)$ (normalized)	$W(i)$	$W(i)$ (normalized)
acount	0.102	0.171	0.064	0.313
degree	0.104	0.174	0.084	0.412
pr	0.062	0.104	0.048	0.238
cls	0.119	0.200	0.090	0.445
btw	0.079	0.132	0.054	0.267

IX.   CONCLUSIONS

A social network of co-occurrence based on Reuters-21578 corpus is constructed. Complex Network techniques are used to analyze the network. It is found to be a small-world network with power-law degree distribution. The distribution of the component size is governed by a power law.

We reduce the network by taking vertices with minimum degree of 10, where the selection of 10 is arbitrary and obtained a network with 49 strongly connected persons. This reduction enables us to interpret the results of a community detection algorithm for a large network with unknown community structure. The algorithm produced 4 communities that we can give a meaningful interpretation.

When the network is reduced by removing the edges whose weight is less than 10, where the selection of 10 is also arbitrary, the network become disconnected. Interestingly, it becomes almost a star-connected network where Ronald Reagan is in the center as a sign of his dominance. Removal of him would break the already disconnected reduced
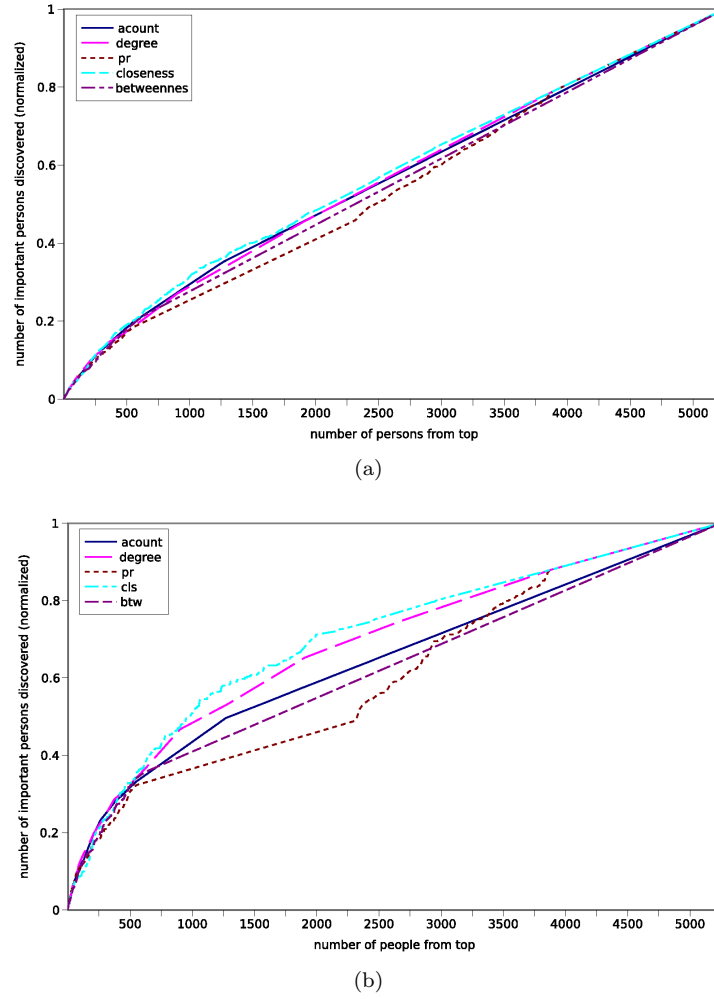
(a)

(b)

FIG. 4: (Color online) Cumulative performance of algorithms for (a) $H(i)$, (b) $W(i)$.

network and transform to a network where only a few vertices are connected.

The articles have time information. Using this, we have obtained time dependency of co-occurrence. Some people such as Ronald Reagan have activity, consistently. On the other hand, some people have limited time presence due to some particular events.

We try to estimate the importance of the people based on the co-occurrence network of 1987 and compare it with another importance metric of 2007 based on representation in the English Wikipedia, giving 20 years of time difference in between. The time past can make a reliable judgment of who was *not* already important in 1987. We present a comparative performance analysis for five common centrality measures for our network; article count, degree, PageRank, betweennes and closeness. We found low to medium normalized Spearman correlations between the two metrics. These were in our expected range and in this sense an important contribution of this paper is in providing a means of comparing the relative performances of the algorithms. We found PageRank to be the worst of the five, while closeness was the best performing algorithm. We have discovered an anomaly with PageRank which to some extent can explain why it was performing badly. We believe this may be important as it points out to a pitfall using PageRank in such a network. It is possible to use and combine results from Wikipedias of different languages, as well as different reference sources which we see as a future work.

Acknowledgments

[1] M. E. J. Newman, *SIAM Review* **45** 2, p. 167 (2003).
[2] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, *Physica A* **311**, p. 590 (2002).
[3] J. D. Watts and S. H. Strogatz, *Nature* **393**, p. 440 (1998).
[4] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, *Nature* **411**, p. 907 (2001).
[5] M. Kirlidog and H. Bingol, *5th International IT in Regional Areas Conference*, Central Queensland, Australia, p. 40 (2003).
[6] V. E. Krebs, *First Monday* **7** 4 (2002).
[7] D. D. Lewis, Reuters-21578 corpus, Distribution 1.0. available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.
[8] A. Özgür and H. Bingol, *LNCS* **3280**, p. 688 (2004).
[9] A. Özgür, Reuters-21578 Social Network, available at http://soslab.cmpe.boun.edu.tr.
[10] D. M. Bikel, S. Miller, R. Schwartz, and R. M. Weischedel, *In Proceedings of the 5th Conference on Applied Natural Language Processing*, p. 194 (1997).
[11] V. Batagelj, A. Mrvar, *Connections* **21** 2 (1998).
[12] M. Girvan and M. Newman, *PNAS USA* **99**, p. 7821 (2002).
[13] Iran-Contra Affair at wikipedia, http://en.wikipedia.org/wiki/Iran-Contra_Affair.
[14] Wikipedia, the free encyclopedia. http://en.wikipedia.org/
[15] L. C. Freeman, *Social Networks* **1**, p. 215 (1979).
[16] L. Page, S. Brin, R. Motwani, T. Winograd, Technical Report, Stanford Digital Library Technologies Project (1998).