Reconstructing History of Social Network Evolution Using Web Search Engines

Jin Akaishi[1,6] Hiroki Sayama[1,2,3,4] Shelley D. Dionne[1,4,5] Xiujian Chen[1,5]
Alka Gupta[1,4,5] Chanyu Hao[1,4,5] Andra Serban[1,4,5]
Benjamin James Bush[1,3] Hadassah J. Head[1,3] Francis J. Yammarino[1,4,5]

[1] Collective Dynamics of Complex Systems Research Group
[2] Department of Bioengineering
[3] Department of Systems Science and Industrial Engineering
[4] Center for Leadership Studies
[5] School of Management
Binghamton University, State University of New York
P.O. Box 6000, Binghamton, NY 13902-6000, USA

[6] Kumamoto National College of Technology
2627 Hirayamashinmachi, Yatsushiro, Kumamoto, Japan

{jakaishi,sayama,sdionne,xichen,agupta1,chao2,aserban1,
bbush2,hhead1,fjyammo}@binghamton.edu

Abstract. We propose a simple web search engine based method for collecting approximated historical data of temporally changing social adaptive networks, which are rather difficult to obtain experimentally in conventional research methods. In the proposed method, a search query string is combined with additional keywords that specify inclusion/exclusion of specific years to limit the search results to a particular time point. Using the proposed method, we reconstructed the temporal evolution of a social network from 2005 to 2009 of 93 individuals who are important in the US economy. We measured centralities of those individuals for every year and found several illustrative cases where the temporal change of centrality of an individual correctly captured the actual events that are related to him/her over this time period. These results indicate the effectiveness of the proposed method. Limitations and future directions of research are discussed.

Keywords: Social networks, adaptive networks, network evolution, centrality, data collection, web search engines.

1 Introduction

The importance of temporal dynamics of network topology and their coupling with node/link state dynamics has been increasingly recognized in network science communities [1,2]. However, it is generally difficult to experimentally obtain large-scale data of real-world social network evolution over time [3,4]. The exceptions to this are some well-studied electronic data sets, such as citations in scientific publications and friendship networks in social media (e.g., Facebook and YouTube), which have been causing a concentration of network analysis research on these limited sets of networks.

Recently, Lee et al. [5] proposed a new web search engine based data collection method by which a researcher can easily reconstruct social networks of any kind by simply using the number of Google search results (hits) for two names as the link weight between them. They demonstrated the effectiveness of this method by applying it to the social network reconstruction for the 109th US Senate members. They also considered the temporal change of this network over several months in late 2006. However, the network "snapshot" data had to be acquired during the time period under investigation, so that the entire data collection process required that data be taken over the course of several months. A remaining open question is how one could use web searches to reconstruct the history of social network evolution *retrospectively*.

We propose a similar web search engine based method for collecting approximated historical data of temporarily changing social adaptive networks by adding to a search query string other keywords that specify the inclusion/exclusion of specific years to limit the search results to a particular time point. For example, one can selectively search for results that are most likely relevant only to year 2006 by adding "2006", "-2007", "-2008", "-2009" and "-2010" to the search query string. We implemented a prototype of this data collection method using the Google AJAX Search API, and applied it to the reconstruction of network history from 2005 to 2009 for 93 individuals who are important in the US economy and industries. In this Work-In-Progress paper, we report the procedure of the proposed method and preliminary results obtained using it, as well as its limitations and future directions of research.

2 Method

The data collection method we propose takes as an input a list of keywords to be searched. Keywords can be of any kind, but in what follows, we focus on the names of the people we wish to include in our social network reconstruction. To each name we add some additional personally indefinable information (in this case, the person's affiliation). This helps reduce some of the errors associated with the fact that many people often share the same name.

Search queries are generated from this list as follows:

1. Two entries (corresponding to 2 people) are chosen from the list described above.
2. A year is chosen (e.g., 2007). The search query will be designed to examine the nature of the relationship that existed between the two people during the chosen year.
3. To eliminate the influence of search results corresponding to documents created in years after the chosen year, we compose a series of partial search queries which exclude the unwanted years. This is done in Google by placing

a negative symbol directly to the left of the unwanted year. That is, if the chosen year is 2007, then the partial search queries -2008, -2009, and -2010 must be included in the search query. Since documents often discuss events which occurred in previous years, we did not add keywords to exclude years prior to the chosen year.

4. Our completed search query is composed by putting together the elements described in steps 1, 2 and 3. For example, to examine the relationship that existed between Warren Buffet and Alan Greenspan in 2007, we used the following search query:

   “Alan Greenspan” “Federal Reserve” “Warren Buffett” “Berkshire
   Hathaway” “2007” -“2008” -“2009” -“2010”

Search queries were generated for all possible pairs from the list of people under study for every year. Each search query was then used to perform a search using the Google AJAX Search API. The code was implemented in Java. The number of search results obtained from each search was recorded and used as the weight, *w*, of the corresponding link of the social network in that year. Note that link weights in this social network are symmetric. To capture the potential asymmetry in social relationships, we calculate the asymmetric weight, $\hat{w}$, of each directed link between keywords *i* and *j* at year *t* as follows:

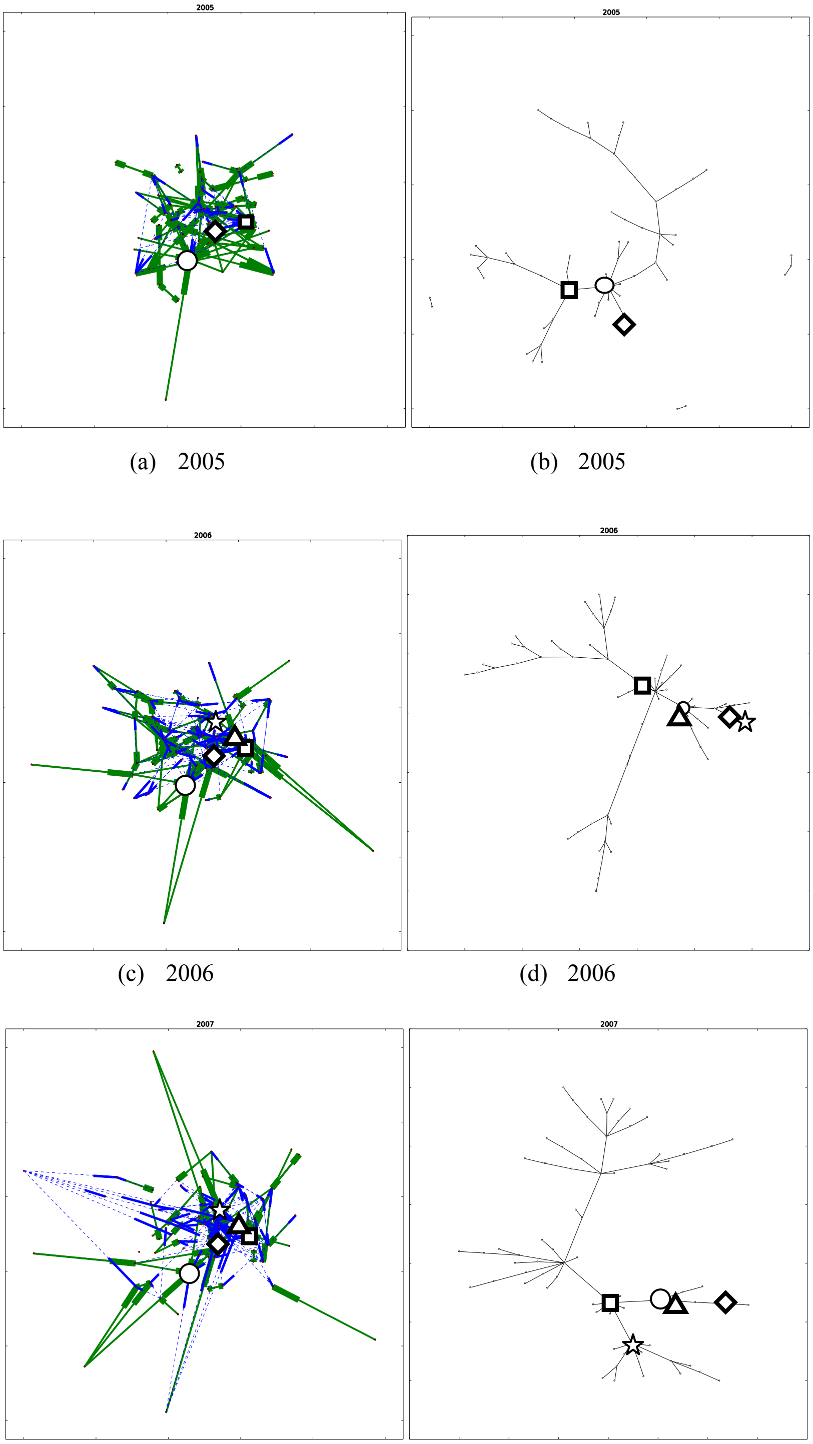
$$\hat{w}(i,j,t) = \frac{w(i,j,t)}{\sum_{k \neq i} w(i,k,t)}$$

For simplicity, we will henceforth refer to “directed networks,” “directed links,” and “directed weights” simply as “networks”, “links,” and “weights,” respectively.

3 Preliminary Results

We conducted a preliminary experiment to test the proposed method. We made a list of 93 people who have played major roles in the US economy recently. Using this list, we constructed a social network for each year from 2005 to 2009 as described in section 2. We chose these years in order to observe relational changes surrounding the 2008 economic crisis. Data collection was done using a single laptop over 15 hours.

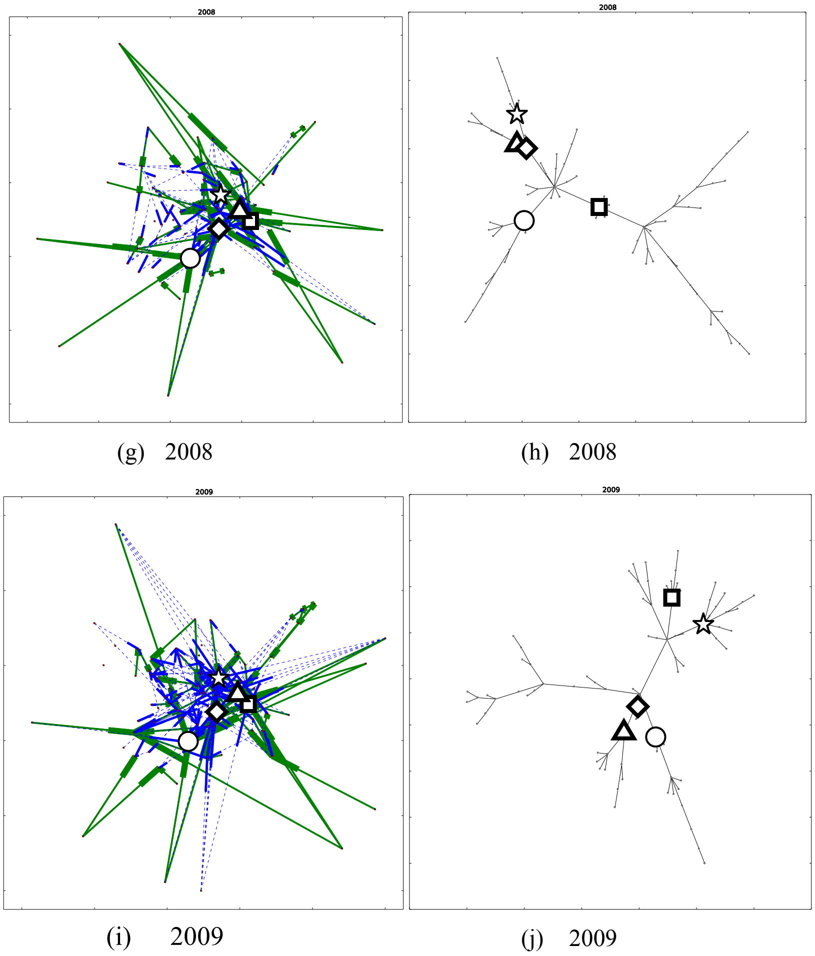
Figures 1 and 2 illustrate the results of the experiment. Figure 1 (a), (c), (e), (g) and (i) are the network visualizations corresponding to the data obtained for the years 2005, 2006, 2007, 2008 and 2009, respectively. Dotted lines indicate weak connections, e.g. links whose weights are between 0.1 and 0.2. Solid lines indicate strong connections, e.g. links whose weights are greater than 0.2. The same node positions are used in all network visualizations so that the changes in link structure will be visible. Figure 1 (b), (d), (f), (h) and (j) are maximum spanning trees which are created based on the same data as the network visualizations. Symbols in the figures show the positions of 5 individuals of interest: Squares, triangles, circles, stars and dimonds indicate Warren Buffett, Timothy Geithner, Alan Greenspan, Lloyd Blankfein and Henry Paulson, respectively.



(a) 2005 (b) 2005

(c) 2006 (d) 2006

(e) 2007 (f) 2007



(g) 2008 (h) 2008

(i) 2009 (j) 2009

Fig. 1. (a), (c), (e), (g) and (i): Network visualizations corresponding to the data obtained for the years 2005, 2006, 2007, 2008 and 2009, respectively. Dotted lines indicate weak connections. Solid lines indicate strong connections. All the network visualizations use the same node positions. (b), (d), (f), (h) and (j): Maximum spanning trees which are created based on the same data as digraphs. Symbols on the each graph show the positions of 5 individuals of interest: Squares, triangles, circles, stars and dimonds indicate Warren Buffett, Timothy Geithner, Alan Greenspan, Lloyd Blankfein and Henry Paulson, respectively.

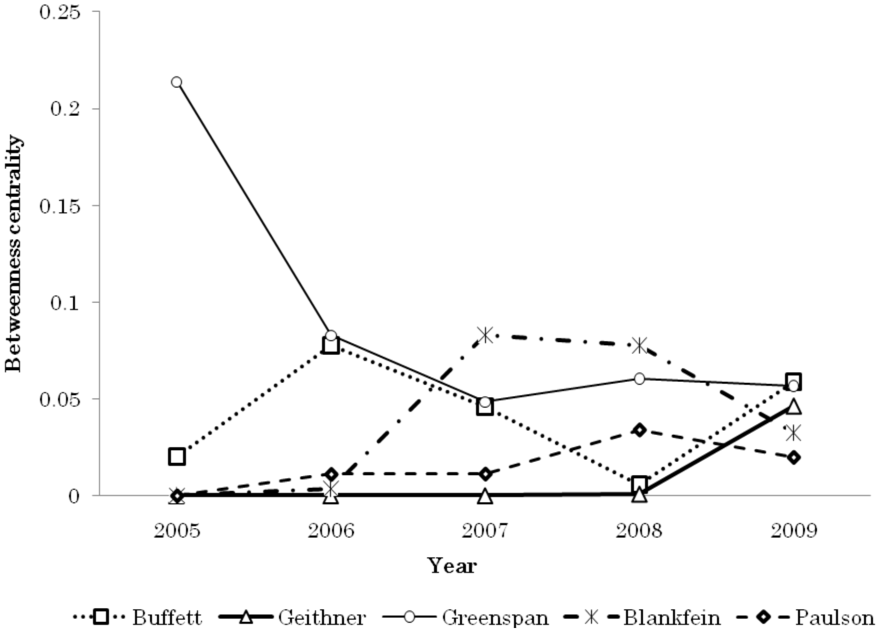


Fig. 2. Betweenness centralities of 5 individuals of interest from 2005 to 2009.

Figure 2 shows temporal changes in betweenness centrality of the 5 well known individuals from 2005 to 2009. The betweenness centrality in Fig. 2 is normalized for comparison so that the sum of the betweenness centralities of all nodes is 1. These plots correctly reflect several actual events that happened in the US ecomony. A straightforward example is that the centrality of Alan Greenspan, who was the chairman of the Federal Reserve until 2006, drastically decreased over time. In contrast, the centrality of Timothy Geithner, who became the US Treasury Secretary in 2009, increased in 2009.

An interesting observation on the relationship between Lloyd Blankfein, CEO of Goldman Sachs, and Henry Paulson, US Treasury Secretary and former CEO of Goldman Sachs, is that their closeness in this social network varies from time to time. Their closeness in 2006 (Fig. 1 (d)) is apparently because Blankfein took Paulson's position as Goldman Sachs CEO in that year. They were apart in the following year (Fig. 1 (f)), but again came close to each other in 2008 when the economic crisis occurred (Fig. 1(h)). Although it was not widely known, Goldman Sachs was AIG's largest trading partner in 2008. When Blankfein noticed that AIG was having very severe liquidity problems, he called Paulson for help. Paulson's office calendar at the Treasury, obtained by the New York Times through a Freedom of Information request, revealed that he spoke to Blankfein two dozen times during the September week when the Treasury bailed out AIG. That was "far more frequently" than Paulson talked to any other Wall Street executive [6,7]. Our network visualization for 2008 correctly captures this strong relationship between Blankfein and Paulson.

These correspondences between the history of those individuals and the changes observed in the figures suggest that our method is capable, to some extent, of illuminating changes that have occurred in real-world social networks.

4 Conclusion

In this paper, we have proposed a new method for acquiring historical social network data retrospectively using web search engines. Data were collected by recording the number of Google search results yielded by carefully crafted search queries. To test our proposed method, we conducted a preliminary experiment in which we reconstructed a social network of 93 important figures in the US economy. Five annual snapshots of the network evolution were generated for years ranging from 2005 to 2009. Temporal changes in network topology and node centrality measure reflected several real-world events, such as shifts of power/influence and temporary formation of strong relationships. These results demonstrate the potential of our proposed method for examining changes that have occurred in real-world social networks.

Our method has several limitations. First, our method relies solely on simple Google searches using heuristically crafted search queries, and therefore its validity is yet to be fully examined and established. Additional filtering using more contextual information may help improve the quality of search results. Our method also has a computational limitation, since the number of search queries required to construct a social network grows quadratically as the number of keywords increases. Yet another limitation is the error in search results. Multiple attempts of an identical Google search query sometimes produce very different results: the number of search hits might vary in orders of magnitude. In order to minimize this type of error, we need to collect data multiple times (e.g., using multiple PCs) and aggregate them.

In future studies, we plan to conduct a more rigorous validation of our data collection method and to develop more advanced temporal filtering techniques using textual contents of the documents returned by web search engines.

References

1. Gross, T., Sayama, H., eds.: Adaptive Networks. Springer, Heidelberg (2009)
2. Braha, D., Bar-yam, Y., : From Centrality to Temporary Fame: Dynamic Centrality in Complex Networks. Complexity Vol.12/No. 2, 59--63 (2006)
3. Doreian, P.; Stokman, F. N., (Eds.) : Evolution of Social Networks. Gordon and Breach, New York (1997)
4. Wasserman, S., Faust, K., : Social Network Analysis. Cambridge University Press: Cambridge (1999)
5. Lee, S.-H., Kim, P.-J., Ahn, Y.-Y., Jeong, H.: Googling social interactions: Web search engine based social network construction. PLoS ONE e11233 (2010).
6. Morgenson, G., : Behind Insurer's Crisis, Blind Eye to a Web of Risk. New York Times, September 27, A1 (2008) http://www.nytimes.com/2008/09/28/business/28melt.html?fta=y

7. Clark, A., : How close are Goldman Sachs's connections with the US treasury?. the U.K. Guardian, August 10, (2009) http://www.guardian.co.uk/business/andrew-clark-on-america/2009/aug/10/goldman-sachs-aig-treasury