# Can Network Analysis Reveal Importance? Degree Centrality and Leaders in the EU Integration Process

Marten Düring(✉)

CVCE, Luxembourg, Luxembourg
marten.duering@cvce.eu

**Abstract.** This paper describes ongoing work on the potential of simple centrality algorithms for the robust and low-cost exploration of non-curated text corpora. More specifically, this paper studies (1) a network of historical personalities created from co-occurrences in historical photographs and (2) a network created from co-occurrences of names in Wikipedia pages with the goal to accurately identify outstanding personalities in the history of European integration even within flawed datasets. In both cases Degree centrality emerges as a viable method to detect leading personalities.

**Keywords:** Historical network research · Wikipedia · Photos · Digital history

## 1 Introduction

Most scholars in the historical network research domain today work with carefully built datasets, some of which take years to complete. There is general scepticism towards automatically generated data [9]. Such data is however attractive given both the volume of texts which can be processed and the speed with which this can be done. Most scholars who specialise in text analytics work with tools far beyond the skill set of most humanists. The question therefore is: How to make the most of (in this case) Social Network Analysis without advanced and in many respects costly methods? To which extent can simple tools yield output which requires only basic technical skills, can be trusted and does not need to be subjected to manual verification? This is especially relevant for selection processes when facing large document collections and helps to reduce the number of potentially relevant nodes (e.g. persons, documents, etc.). The approach for this experiment is therefore consciously both naïve and simplistic. The goal is to find out whether easily obtainable yet by no means authoritative datasets (more on this below) can still be used to identify key actors. To which extent can photographs and Wikipedia pages be understood as proxies for real social relations and interactions? Any lessons learned in such a controlled environment will be beneficial for the analysis of similar approaches to unknown datasets.

Centrality measures developed in social network analysis are applied to two strongly biased network datasets and used to explore to which extent centrality measures are capable of identifying important actors in the history of European integration.


L.M. Aiello and D. McFarland (Eds.): SocInfo 2014 Workshops, LNCS 8852, pp. 314–318, 2015.
DOI: 10.1007/978-3-319-15168-7_39

There are many ways to define importance based on very different metrics and any definition will be based on more specific premises. Any understanding of importance attributed to individuals is the result of human-made, reversable selections which single out some while hiding others, some of which get rediscovered, some remain forgotten. But without doubt, some people left a greater mark on history than others. More specifically, importance in this context describes 1) having held high ranking offices in European institutions, 2) subjective judgment and best knowledge of the domain which leads me to attribute importance to Charles de Gaulle but not to Emilio Colombo, to Pierre Werner but not to Edward Heath (choices documented below). Any such list must be necessarily fuzzy; there can be no such thing as a universal, ranked list of important leaders and it would not be of much interest for historians anyways. Such choices are subject to debate, however a comparison between the persons with the highest and lowest scores does suggest plausibility to the overall findings of this paper. There are, I am sure, ways to find and compute abstract notions of importance. In this paper however I seek to understand to which extent centrality metrics applied to two imperfect datasets can still yield subjectively relevant results.

## 2 Literature Review

Historians in all subdisciplines have adapted methods and theories developed in Social Network Analysis in very different ways for several decades, in recent years interest as risen significantly [9]. The literature on social network extraction from photographs is still sparse and e.g. concerned with inferences of social ties based on celebrity photos [12] and race relations [1]. [2] find that (artificial) networks remain resilient to minor distortions. Wikipedia and the related DBpedia project have been used for a very large number of related research projects; only a subset of which can be listed here. Social Network Analysis has most often been used to study user interaction in Wikipedia [3]. [10] use Wikipedia for entity disambiguation, [11] use it to extract tripartite networks. More relevant to this topic is work on network extraction from texts in general. [6] note the importance of enriched and cleaned data and propose a meta-matrix approach for the detection of related concepts in texts, based on which they infer social ties. [4] apply sentiment analysis to extract positive and negative ties from biographies of 19$^{th}$ century Dutch socialists. [5] Infer social ties from geographic coincidences. Earlier work by the author analyzed covert historical networks of help for Jewish refugees during the Holocaust [8]. Nodes and edges were manually coded from text and the performance of centrality for a list of actors strongly involved in the respective networks was evaluated. In these networks, which were collected with great care, on average 67 percent of all strongly involved actors also appeared in a group of actors with the top 20 percent highest centrality scores. Degree centrality outperformed all other centrality measures.

## 3 Data Collection

Based on their collection of photographs, Centre virtuel de la connaissance sur l'Europe (CVCE) compiled a list of 468 personalities who are associated with the history of European integration. CVCE researches and tells the story of European integration based on corpora of primary sources, each of which explores a different aspect of this highly complex process. We assume that they have obtained at least one portrait photograph of all individuals they have considered to be of outstanding importance in this context. CVCE's focus on primary sources also means that every photograph with more than one person represents an historical event of some significance and by extension, at least some of the persons in the photograph must have played a significant role in the pictured event. Based on these considerations we can state that the list of 468 persons contains nearly all major actors in the process of European and that the list will also contain people of lesser importance who were photographed alongside others.

Wikipedia is the single most complete dataset for such information which is freely available and is an important source for information for many. For these reasons we can attribute it some significance as an indicator of a persons perceived status. A Python script was used to download the English Wikipedia page of each of the 468 persons. References to any of the other 467 persons in each personal page constitute an edge between the two. This applies to occurrences of Wikipedia-IDs (e.g. « Jean-Luc_Dehaene ») in the descriptive text and in the structured text (e.g. « Prime Ministers of Belgium »), additional occurrences increase an edge's weight. The data was cleaned using Google Refine and checked manually. This results in a graph containing 461 nodes and 5288 edges. Removal of isolates and self-loops reduced the graph to 369 nodes and 4628 edges with an overall low density of 0.068. This graph was used for the following computations of centrality scores.

The second graph was downloaded from histoGraph [13], a tool developed by the EC FP7-funded research project CUbRIK which brought together scholars in multimedia search and human-machine interaction. histoGraph combines automatic and crowd-based face recognition and identification. The tool currently contains a social network of 222 of the 468 individuals in CVCE's photograph collection. A weighted edge is created for each co-occurrence of two persons in a photograph. This yields 371 edges and an even lower overall undirected density of 0.024. Density describes the ratio of existing edges in a network to the number of possible edges.

## 4 Analysis

Commonly used centrality measures for both graphs were computed using Gephi's SNA Metrics Plugin. The following measures were selected: Degree, Betweenness, Closeness, Eigenvector, PageRank and Clustering Coefficient. For each measure the 25 highest scoring persons in both datasets were compared to get a first sense for their performance. Only Degree centrality scores came somewhat close to the expected results. Table 1 lists the highest scoring persons for degree centrality, which represents the number of ties a node has. An asterisk indicates importance attributed by me. It lists persons which fit the vague definition of importance such as Francois Mitterrand

or Konrad Adenauer. In the Wikipedia network, among the 25 highest ranking persons 20 can be considered important and 14 in the histoGraph network. Others like Alois Mock seem to have a stronger profile in their respective home states. Still others such as US or Russian presidents can be considered important, albeit not primarily in the context of Europe. It would be misleading to filter the latter out based on these or other distinctions. Instead I chose to check all of the lower ranking persons in both networks for importance. Only three notable personalities in the Wikipedia network scored rather low in the histoGraph network (in brackets their degree): *Charles de Gaulle* (4), *Francois Mitterrand* (3), *Alcide de Gaspari* (4). In the Wikipedia network *Pierre Werner* (5) and *Jean Monnet* (5) have surprisingly low degrees. The complete dataset including all centrality scores is available online [7].

**Table 1.** Highest degree scores in both networks, cut-offs at 61 for the Wikipedia and 5 for the histoGraph network. Subjectively important persons are highlighted with an asterisk.

| Rank | Wikipedia network | | histoGraph network | |
|---|---|---|---|---|
| | Name | Degree | Name | Degree |
| 1 | Francois Mitterrand* | 129 | Konrad Adenauer* | 27 |
| 2 | Helmut Kohl* | 114 | Robert Schuman* | 21 |
| 3 | Walter Hallstein* | 108 | Margaret Thatcher* | 17 |
| 4 | Konrad Adenauer* | 106 | Walter Hallstein* | 15 |
| 5 | Felipe Gonzalez* | 103 | Paul-Henri Spaak* | 14 |
| 6 | Helmut Schmidt* | 101 | Pierre Werner* | 13 |
| 7 | Charles de Gaulle* | 97 | Helmut Kohl* | 12 |
| 8 | Alcide De Gasperi* | 95 | Joseph Bech | 11 |
| 9 | Margaret Thatcher* | 93 | Jean Monnet* | 11 |
| 10 | Winston Churchill* | 92 | Franz Vranitzky | 10 |
| 11 | Giulio Andreotti* | 86 | Amintore Fanfani | 9 |
| 12 | Aldo Moro | 85 | Jacques Santer* | 9 |
| 13 | George Marshall* | 85 | Willy Brandt* | 8 |
| 14 | Robert Schuman* | 85 | Helmut Schmidt* | 8 |
| 15 | Bettino Craxi | 82 | Hannelore Kohl | 8 |
| 16 | Willy Brandt* | 80 | Paul Finet | 8 |
| 17 | Anibal Cavaco Silva | 77 | Antoine Pinay | 8 |
| 18 | Guy Mollet | 75 | Klaus Hänsch | 7 |
| 19 | John F. Kennedy* | 75 | Valery Giscard d'Estaing* | 6 |
| 20 | Valery Giscard d'Estaing* | 75 | Franz Etzel | 6 |
| 21 | Jacques Delors* | 74 | Alois Mock | 6 |
| 22 | Jean-Claude Juncker* | 73 | Gaetano Martino | 6 |
| 23 | Jimmy Carter* | 73 | Georges Pompidou* | 6 |
| 24 | Ronald Reagan* | 73 | Paul Reynaud | 6 |
| 25 | Alain Poher | 72 | Leo Tindemans* | 6 |
| 26 | Dwight D. Eisenhower* | 72 | Winston Churchill* | 6 |
| 27 | Edward Heath | 70 | Hans-Dietrich Genscher | 6 |
| 28 | Jacques Chirac* | 69 | Gaston Thorn | 6 |
| 29 | Joseph Stalin* | 69 | Ronald Wilson Reagan* | 6 |
| 30 | Georges Pompidou* | 68 | Yasuhiro Nakasone | 6 |
| 31 | Paul-Henri Spaak* | 67 | Enzo Giacchero | 5 |
| 32 | Richard Nixon* | 67 | Thomas Klestil | 5 |
| 33 | Tony Blair | 67 | Giulio Andreotti* | 5 |
| 34 | Erich Honecker | 66 | | |
| 35 | Harry S. Truman* | 66 | | |
| 36 | Emilio Colombo | 65 | | |
| 37 | George H. W. Bush* | 65 | | |
| 38 | Harold Wilson* | 65 | | |
| 39 | Leo Tindemans* | 65 | | |
| 40 | Lester B. Pearson | 65 | | |
| 41 | Urho Kekkonen | 65 | | |
| 42 | Harold Macmillan* | 63 | | |
| 43 | John Foster Dulles | 63 | | |
| 44 | Vyacheslav Molotov | 63 | | |
| 45 | Andrei Gromyko | 62 | | |
| 46 | Ernest Bevin | 61 | | |
| 47 | Fidel Castro* | 61 | | |
| 48 | Jacques Santer* | 61 | | |
| 49 | Romano Prodi* | 61 | | |

This suggests that we can safely expect to find a large majority of subjectively important persons among the highest ranking degree scores in both networks.

## 5 Future Work

At this stage I treated edges in both networks as undirected. Future work will consider the directionality of these co-occurrences since it does make a difference whether for example King Albert II is mentioned on Dehaene's page or Dehaene on King Albert's. The surprisingly poor performance of other centrality measures to detect importance must not mean that they are useless. At this stage it remains open whether high scores for less known persons indicate indeed some kind of influence or must be treated as artifacts.

## References

1. Berry, B.: Friends for Better or for Worse: Interracial friendship in the United States as seen through wedding party photos. Demography **43**(3), 491–510 (2006)
2. Borgatti, S.P., et al.: On the robustness of centrality measures under conditions of imperfect data. Social Networks **28**(2), 124–136 (2006)
3. Brandes, U., et al.: Network analysis of collaboration structure in Wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, pp. 731–740. ACM, New York (2009)
4. Van de Camp, M., van den Bosch, A.: The socialist network. Decision Support Systems **53**(4), 761–769 (2012)
5. Crandall, D.J., et al.: Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences **107**(52), 22436–22441 (2010)
6. Diesner, J., Carley, K.: Exploration of communication networks from the enron email corpus. In: Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005, pp. 3–14 (2005)
7. Düring, M.: Appendix for the HistoInformatics 2014 paper submission, https://dl.dropboxusercontent.com/u/132496/Permanent/Wikipedia-histoGraph-centrality-scores-histoInformatics2014.xlsx
8. Düring, M.: How reliable are centrality measures for data collected from fragmentary and heterogeneous historical sources? A case study. In: Proceedings of The Connected Past Conference. Oxford Publishing, Oxford (forthcoming)
9. Düring, M., Eumann, U.: Historische Netzwerkforschung. Ein neuer Ansatz in den Geschichtswissenschaften. Geschichte und Gesellschaft **39**, 369–390 (2013)
10. Gattani, A., et al.: Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-based Approach. Proc. VLDB Endow. **6**(11), 1126–1137 (2013)
11. Nazir, F., Takeda, H.: Extraction and analysis of tripartite relationships from Wikipedia. In: 2008 IEEE International Symposium on Technology and Society, ISTAS 2008, pp. 1–13 (2008)
12. Ravid, G., Currid-Halkett, E.: The social structure of celebrity: an empirical network analysis of an elite population. Celebrity Studies **4**(2), 182–201 (2013)
13. Wieneke, L., et al.: histoGraph – A Visualization Tool for Collaborative Analysis of Historical Social Networks from Multimedia Collections. In: Proceedings of 18th International Conference Information Visualisation (IV), 2014 Conference, Paris, France (2014)