The Content Structure of Intelligence Reports

Kimmo Elo
University of Turku
Turku, Finland

Abstract

Despite its close connection to many of the methodological questions and problems related to uncloaking hidden structures or to analyzing network dynamics tackled by network analysts in terrorism or crime studies, researchers on Cold War intelligence have shown limited interest in network analysis. Although there might be material-related reasons for that, the resistance is to a great extent caused by the unfamiliarity with the method itself. Following the idea that how one looks at the material determines what they see, this article will evidence how social network analysis could be applied to historical sources in order to extract, analyze and visualize new knowledge. The analysis will illustrate the capability of SNA-based keyword co-occurrence network analysis and visualizations for uncovering and identifying the corpus’ structural properties, detecting the thematic backbone of the report corpus, and for analyzing network dynamics. The material used in the analysis consists of reports on Nordic affairs produced in 1975-1989 by the East German foreign intelligence service. Besides keyword co-occurrence network analysis and visualizations, the article shows how community detection techniques can be used to extract thematic backbones in a report corpus. A critical assessment of the results obtained by SNA techniques in their historical context confirms their validity and reliability. More generally, the results showed that combining SNA with methods of content analysis offers a promising perspective for developing new research methods for the analysis of the social network of language capable of tackling and extracting contextual and semantic relationships in networks based on textual data. Furthermore, the SNA-based approach to textual networks could open up new perspectives for exploratory historical research with the view to finding new foci for research and comprehending new research hypothesis and questions.

Keywords: SNA, Co-occurrence networks, Content analysis, Network communities, Visualizations, East German foreign intelligence, Intelligence reports, Cold War

Authors

Kimmo Elo, is an adjunct professor and senior lecturer in the Institute for Political Science and Contemporary History at the University of Turku in Turku, Finland and The Institute for German Studies at the Åbo Akademi University in Turku, Finland. His research interest include Historical network research, German politics and history, Intelligence studies, Cold war history, European integration and Digital Humanities.

Correspondence concerning this work should be addressed to Kimmo Elo, Department of Political Science and Contemporary History, FIN-20014 University of Turku, Turku, Finland. Email: kimmo.elo@utu.fi.

DOI: http://dx.doi.org/10.17266/35.1.2
1. Introduction

The very essence of all intelligence services is related to their networking capacity. A closer look at the so-called intelligence cycle, the cyclical process consisting of information-gathering (single-source collection), exploitation, (all-source) analysis and dissemination, quickly reveals how fundamentally important networking capabilities and networks are for all intelligence services. First, each intelligence service attempts to build a functioning, dense network of human (HUMINT\(^3\)) or non-human (SIGINT\(^4\)) sources providing the service with raw data. In intelligence analysis, analysts combine information from the collected raw data with previous knowledge and information from other sources in thematic reports. From this perspective an intelligence report consists of logically connected pieces of information, i.e. an information network. Finally, these reports are disseminated to a small or large network of recipients at various levels of governmental administration.

Taking into account the importance of intelligence during the Cold War and the centrality of networks and networking (both human and non-human) for intelligence services, the limited interest in network analysis among Cold War studies is somewhat surprising. One reason might be that Cold War studies are still dominated by historians who are neither primarily interested in intelligence studies\(^5\) nor familiar with network analysis as a research method\(^6\). Conversely, recent research in terrorism or crime studies has discussed empirical, methodological and theoretical questions and problems in relation to uncovering hidden structures or analyzing network dynamics\(^7\), which are also relevant for intelligence studies. But, outside the humanities, the opposite holds true as network studies have shown only modest interest in historical sources. There may be a rational explanation for this lack of interest. If empirical data is used just for algorithm testing, the use of materials requiring in-depth knowledge of source criticism and time-consuming preparation simply makes no sense.

Despite its close empirical connection with the Cold War period, this article intends to contribute to methodological and theoretical discussions in intelligence studies. Following the idea that research methods have a strong impact on what can be achieved, this article will show how network analysis could be applied to historical sources in order to extract and visualize new knowledge. The article focuses on two questions. First, how can social network analysis (SNA) be applied in order to study the characteristics of co-occurrence networks of keywords summarizing the content of intelligence reports? Second, how can the meta-data of intelligence reports be used for tackling characteristics and dynamics of the report corpus? The analysis will also illustrate the capability of clustering techniques for unclouking and identifying the corpus’ structural properties, detecting the thematic backbone of the report corpus, and for analyzing network dynamics.

The structure of this article is as follows. Section two discusses the research method and material used. On this basis, the article will introduce a technique combining co-occurrence analysis and a method for identifying the key concepts in a document collection functioning as junctions of different documents. In regard to material, the posthumous data from the HV A archive available for research will be introduced. Section three consists of analysis and visualizations focusing on content structures, dynamics and thematic clusters in the keyword co-occurrence networks. The article will conclude with a critical assessment of the most important findings in the broader context of Digital Humanities.

2. Method and Material

2.1 Data Analysis and Visualization Methodology

The idea of unclouking hidden structures from texts with the help of graphical tools and representation is not new. Since the late 1990s, humanists and social scientists have shown an increasing amount of interest on network analysis, a development, which has resulted in the emergence of new concepts and methods\(^8\), including text mining, semantic network analysis and content analysis, which are currently widely used by humanists focusing on structure-oriented analysis of large text materials.

The concept of co-occurrence networks is a

---

\(^2\) Herman, 2001, 79; Bruce & George, 2008, 2; Walsh, 2011

\(^3\) HUMINT stands for "human intelligence", i.e. for intelligence-gathering by interpersonal contacts.

\(^4\) SIGINT stands for "signal intelligence", i.e. intelligence-gathering by intercepting all kind of communication or electronic signals.

\(^5\) Garthoff, 2004, 21

\(^6\) Although network analysis has been used in a wide range of historical case studies (an up-to-date bibliography of historical network research can be found at: https://www.zotero.org/groups/historical_network_research_bibliography/items [on-line: visited on August 10, 2014]), to our knowledge it has not been applied to historical intelligence studies.


\(^8\) For a recent review, see Schultz-Jones, 2009
powerful method for the graphical representation of potential and existing relations between different concepts, terms or other entries in textual materials. Co-occurrence networks are widely used among scholars interested in content analysis or text mining through which an understanding of the thematic structure of a text corpus is sought. In recent studies, co-occurrence network based approaches have been applied in constructing semantic networks of scientific journals, extracting social, issue or concept networks from large datasets, analyzing customer feedback or assessing free text answers, and in creating semantic summaries of documents.

Notwithstanding its growing popularity among scholars, co-occurrence network analysis has mainly been used to examine and visualize relations between concepts in order to find out the most or the least connected concepts, semantic clusters within a text corpus or changes over time in co-occurrences. Scholars have shown less interest in the network structure itself, its characteristics and network metrics. Although most studies on co-occurrence networks have several graphical representations, the visualizations are rarely discussed, let alone analyzed. In most papers visualization settings (layout used for visualization, layout settings etc.) and their impact on the graphical representations are simply left unexplained. One central reason for this lack of interest in network structure and characteristics is the fact that most co-occurrence analyses have focused on the interconnections of concepts, not their interaction or influence on the network structure as a whole. In other words, in most co-occurrence analyses visualizations serve as graphical representation of the content in the form of paired concepts.

Mastering the shift from the analysis of simple concept interconnections to concept co-occurrence networks requires a methodological bridge-building from co-occurrence analysis to SNA. The analysis and visualizations conducted in this article are based on Paranyushkin’s (2011) recent work on using SNA for identifying meaning circulation in text documents as well as on Feicheng & Yating’s (2014) study of the use of SNA for co-occurrence network analysis of on-line tags. Both papers share the relatively simple, yet powerful understanding that SNA - originally designed for analysis of social or human relationships - could be applied to co-occurrence networks of textual data as well. Although this slightly changes the terminology "actors" are replaced by "concepts" or "keywords" - it does not affect the focus; the emphasis remains on exploring and understanding the structures and internal relationships among concepts / keywords, not just on presenting the connections between them. Both papers also powerfully exemplify how SNA methods can be used to extract and visualize the basic characteristics and the structure of the co-occurrence network. This article seeks to illustrate how network metrics can be used for comparisons between different time periods in order to uncover network dynamics.

The method for generating co-occurrence networks of report keywords is based on the understanding that each report forms a particular thematic space which is summarized in keywords. The underlying assumption is that keywords are used to describe content in a similar way than tags are used for tagging on-line documents. Consequently, if one creates keyword co-occurrence matrices for each report in the report corpus, it can be assumed that thematically comparable reports will have an analogous keyword co-occurrence structure. Moreover, because the East German foreign intelligence service used a standardized set of keywords for report tagging and the keywords were added according to the principle of relevance, we are expecting to identify core keywords constructing the core knowledge of the report corpus, i.e. keyword co-occurrences linking different reports to each other. To keep the order of relevance, the co-occurrence matrix is constructed by pairing the first keyword (which is considered as the most relevant) with all subsequent keywords. Based on this method, a network of keywords and their co-occurrences - G(V, E) - is built, where V is the set of keywords and E is the set of edges (co-occurrences).

2.2 Intelligence Report Data of East German Foreign Intelligence

The East German foreign intelligence service’s main department A (Hauptverwaltung A, abbr. HV A) was formally a part of the East German State Security Service (Ministerium für Staatssicherheit, abbr. MfS, Stasi). Like all intelligence services, the HV A was responsible for running the complete intelligence cycle from information...
gathering to report dissemination according to objectives and guidelines decided by the party and state leadership\(^{16}\), i.e. "torn between its twin skills of collecting information and evaluating it"\(^{17}\). The main task of the HV A was to gather, evaluate and process technical, scientific, military and political information. It was also responsible for compiling reports on political, economic and strategic issues, disseminated not only to responsible state and party organizations inside the GDR, but also to allied services inside the Soviet Empire, most importantly to the Soviet KGB. Put in theoretical terms, the HV A was a huge information system responsible for collecting, storing, analyzing and disseminating information. Its systemic structure was built upon a center-periphery schema: the periphery was responsible for gathering intelligence while the role of the center was to control, direct and supervise activities in the periphery as well as to process the collected information to reports for dissemination\(^{18}\).

The backbone of this information system formed an extensive network of "unofficial collaborators", the IMs (Inoffizielle Mitarbeiter), organized in HUMINT networks operating in all Western European countries. In 1989, approximately 189,000 IMs operated for the Stasi and about 15,000 of them for the HV A\(^{19}\).

In 1990, the archive of the HV A was almost completely destroyed. The remaining materials are maintained by the Agency of the Federal Commissioner for the Stasi records (Der Bundesbeauftragte für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik, abbr. BStU). In total, the Agency maintains approximately 111 kilometers of archived files of which the share of the HV A is just 47 meters\(^{20}\). However, in 1998 experts of the BStU succeeded in decrypting an operational database system of the HV A called SIRA (System der Informationsrecherche der HVA). This system was instigated in the mid-1970s and used to maintain the intelligence cycle, administer undercover operations, and most importantly, to store meta-data of information and reports\(^{21}\). Because SIRA was developed for the maintenance of daily information flows to and from the HV A, it opens a window into the HV A’s daily operational work and provides scholars with operational data related to the intelligence cycle from the perspective of the HV A. The SIRA entries cover the years from 1969 to 1989, but the completeness of the stored information varies a great deal. In addition, SIRA records do not contain original documents, and thus, do not substitute the destroyed archival files. Rather, the SIRA records are comparable to bibliographical entries in a library catalog giving the user information.

---

17 Herman, 2001, 3-4
18 On information systems, see e.g. Avison & Myers, 1995; Alter, 2008; Hyvönen et al., 2008
19 Müller-Enbergs, 2011, 21
20 Müller-Enbergs, 2011, 11-12
21 SIRA was built as a relational database and consists of four main tables (sub-databases) numbered from 11 to 14. Each sub-database is dedicated to a specific domain of intelligence: "Scientific and technical espionage" (sub-database #11), "Problems and operations related to domestic and foreign policies, economy and military politics outside the GDR" (sub-database #12), "Political relations in the operation area" (sub-database #13) and "Counter-intelligence" (sub-database #14). Additionally, administrative information of operations - e.g. supervisor changes, opening of new files - are stored in the sub-database #21. (Konopatzky, 2003; Müller-Enbergs, 2007, 13ff.)
of the collections. In the case of the HV A, however, the collections have been destroyed and only the catalog exists.\textsuperscript{22}

The material used in this article consists of a selected corpus of dissemination (type “SA”\textsuperscript{23}) records on Finnish and Nordic affairs from 1975 to 1989. During the Cold War, the four Nordic countries - Finland, Sweden, Norway and Denmark - formed an interesting geopolitical area in the European north, characterized by both differences and similarities.\textsuperscript{24} As recent studies have shown, the HV A conducted intelligence operations also in the European North. The Nordic countries were an important, but not the central operational area for the HV A and undoubtedly these countries enjoyed a special status for the East German foreign intelligence.\textsuperscript{25} The year 1975 is widely regarded as a turning point in the GDR’s history, but also in European politics. The main reason for this assessment is the Conference on Security and Cooperation in Europe (CSCE) held in Helsinki in August 1975. During the second half of the 1970s, the political consequences of the CSCE resulted in growing tensions within the Soviet empire, including the GDR. The party leadership in the GDR was increasingly concerned about the destabilizing impact of the CSCE

\textsuperscript{22} Researchers cannot directly access the database, but SIRA queries are carried out by BStU according to search criteria defined by the researcher. Since the database is administered in a SQL-based system, complex multi-criteria queries are possible. The results are available in printed form only and BStU charges a small per-page fee (currently 10 cents/page).

\textsuperscript{23} Each SIRA record has an unique ID starting a two-character string (“SE”, “SA” or “SB”) describing the information type, followed with two digits identifying the recording year and five digits from the database counter. Records marked with “SE” (SIRA Eingang) are input records, i.e. meta-data of intelligence gathered by the HV A. Records marked with “SA” (SIRA Ausgang) contain the meta-data of disseminated material (reports, evaluations etc.) the HV A has disseminated to external partners. Finally, records marked with “SB” (SIRA Bestellung) are records storing meta-data for intelligence requests from outside. As an example, a SIRA record with the ID “SA7503201” is a dissemination record (type: SA) stored in 1975 (SA7503201).
on its power and consequently instructed the HV A to conduct continuous evaluations of the situation in Europe. These developments were boosted by Mikhail Gorbachev’s rise to power in the Soviet Union in 1985. Against this background it is worth analyzing whether the HV A reports on Nordic affairs also reflect these tectonic changes in Europe. More generally, the analysis seeks to signal two aspects. First, keywords are a valuable source of knowledge capable of providing valuable information about content structures and dynamics. Second, SNA offers well-suited methods for network analysis revealing valuable information about the relationships within the network. A keyword or tag co-occurrence network should be considered as the embodiment of relevant information about the underlying structure and relationships. Here the devil is the detail: two co-occurrence networks may look very similar, but, at the same time, show significant deviation in network metrics. SNA offers solid methods for capturing the concealed, but important differences.

The report corpus used in the analysis consists of 69 reports. Almost a one third of them, 28 reports, were produced in 1975-1977. Another peak in reporting, a total of 18 reports, occurred in 1984-1986 (see Figure 1). The data preparation was carried out in three steps. First, all paper documents were scanned for optical recognition. Second, each document was processed by Tesseract software for optical character recognition (OCR) and stored in text format. Third, a small Python program for processing the text files in order to recognize different fields, extract field contents, and store the extracted data in a database was developed. Although the results from the OCR processing were of relatively good quality, our program was equipped with some learning and correction functionality. The following information was extracted from the meta-data and stored in the database: 1) original date of the report 2) content keywords 3) country references 4) references to objects (institutions, parties, universities, etc.) and persons. Finally, the keyword co-occurrence matrix was created according the methodology described above. Since the original keywords cross-reference to objects and persons, the keyword co-occurrence matrices were constructed in the following manner: if the special German keyword OBJEKT (object) or NAME (name) appeared in the keyword list, it was removed and the subsequent ”real” keywords were upgraded. Object and person references were added to the end of the keyword vector. This practice should ensure the priority of thematic keywords in the pairing process. The content-related keywords were considered as more important structural properties because their purpose is to summarize the content of the report.

3. Using SNA to Explore Intelligence Reports

The readability and interpretation of all graphical representations of network data depends on the quality of the visualization layout. Although network analysis and visualization software have taken quantum leaps, different programs continue to produce somewhat different results. The following analysis and visualizations are prepared with Gephi. The layout used in visualizations is based on Force Atlas algorithm, which seeks to push the most connected nodes - the so-called hubs - away from each other (in order to visually highlight the different communities in the graph) and to align the nodes connected to a hub in clusters around them. Moreover, the algorithm seeks to position the authorities into the center ground. Thus in the following graphical representations the most important concepts in the report corpus are located in the middle of the graph. The most important layout parameters used for all visualization are listed in Table 1.

The keywords-per-report network visualizes the distribution of different keywords (n=293) along the report corpus (Figure 2). In the graph, each report is presented as a red circle and each keyword as a blue circle. The dense sub-network in the middle indicates that several reports share the same keywords, and thus, it is not possible to identify clear overall allocation of keywords. However, we can isolate some reports, such as SA7603773, SA7504200, SA8890104 and SA8892030

Table 1: Force Atlas layout parameters used for visualizations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repulsion strength</td>
<td>200.0</td>
</tr>
<tr>
<td>Attraction strength</td>
<td>10.0</td>
</tr>
<tr>
<td>Gravity</td>
<td>30.0</td>
</tr>
<tr>
<td>Attraction distribution</td>
<td>True</td>
</tr>
</tbody>
</table>

*a) Distributes the attractive force along outbound links, thus pushing hubs at the periphery and putting authorities more central.
that share only some keywords with the rest of the corpus. This visual observation is confirmed by network statistics: the network density (the actual proportion of ties in a network) is very low ($d=0.012$) and the average degree (the number of links in the network compared to number of nodes) is also low (only 2.222). Regarding individual keywords, the top-5 connected keywords (keywords with the highest degree) are: foreign policy (deg=20), east-west relationship (20), foreign relations (18), economy (17) and the European Community (EC, 16).

Although the keywords-per-document network is useful for identifying core keywords in the corpus, the relationships among the keywords are more interesting in regard to the content structure. Similar to social networks, keyword co-occurrences construct a "social network of language in which the individuals or actors are not the members of a group, but terms [keywords], and the links are the relationships among them". Consequently, the significance of the type of interactions that occur between keywords increases in importance. Both the density of co-occurrences (how often two keywords co-occur) and
the number of keywords that relate to the same keyword are significant when assessing the complexity in the network.\textsuperscript{32}

The next graph visualizes the keyword co-occurrence network (see Figure 3). In order to visually highlight some characteristics of the network structure, visual effects are added to the graph. First, the size of a node is proportional to its degree, i.e. the more connections the node has the larger it appears in the graph. Consequently, keywords co-occurring with many other keywords can easily be identified in the graph. Second, the thickness of an edge is proportional to its weight, i.e. how many times the co-occurrence occurs in the data. Since the pairing method produces singular co-occurrences at report level, a co-occurrence’s weight equals to the number of reports sharing that keyword co-occurrence. Thus, thick edges indicate core keyword co-occurrences in the report corpus. Third, the font size of node labels is proportional to the node’s betweenness centrality. In network theory, centrality indicates a node’s position in the network and can be calculated either relative to a node’s direct neighbors or the whole network. Betweenness, as the term itself indicates, defines centrality by analyzing where a node is placed within the network. Consequently, a node’s betweenness centrality score is computed by taking into consideration the rest of the network and by examining how many times a node sits on the shortest path linking two other nodes to each other. Thus, using betweenness centrality as an attribute for the graph helps us to identify nodes that have a “high probability of occurring on a randomly chosen shortest path between two randomly chosen vertices”.\textsuperscript{33}

In this article, the concept of betweenness centrality is preferred to eigenvector centrality because an attempt is made to uncover thematic pathways in the report corpus. In this respect, keywords functioning as mediators between different clusters / contexts in the report corpus are considered as more significant. A node’s eigenvector betweenness is rather reliant on the ties of the node’s connections, whereas betweenness centrality depends on the node’s capability to act as a connection between two or more nodes that would otherwise remain disconnected. Considering thematic pathways, the latter capability is assumed to be more relevant and therefore betweenness centrality is used to measure and visualize a keyword’s status in the keyword co-occurrence network.

In this article, betweenness centrality is interpreted as an indicator for a concept’s role in the report corpus in a straightforward manner: nodes with high betweenness centrality metric are considered as mediators between different clusters / contexts found in the whole report corpus, thus revealing the variety of contexts the concepts appear in while concepts with a low(er) betweenness centrality metrics are central in a sub-corpus only. In other words, a node with a high degree but a lower betweenness centrality might be an important local hub in one cluster, but less influential within the whole network because it has only few connections with other clusters. In turn, a node with fewer connections (lower degree), but high betweenness centrality metrics is considered to be adjacent to the most reports in the network.\textsuperscript{34}

The overall density of the keyword co-occurrence network is 0.013, which shows only weak connections. The average path length is 3.245, which is relatively short and indicates that each keyword can easily reach another keyword. The clustering coefficient measuring the number of node triangles in a graph is quite low, 0.083. This is due to the size of the network (291 nodes and 582 edges) and the pairing method: suppose one report with keywords A, B and C, resulting in keyword co-occurrences A-B and A-C. Now, a triangle would require another report with the keywords B and C so that either B or C is the first keyword. Actually, the relatively low clustering coefficient metric imply the existence of a set of core keywords forming the thematic backbone of the report corpus.

If this assumption holds true, these keywords should have the highest betweenness centrality values. According to the network data, the three most important keywords include communist party (labelled KP in the graph with the betweenness centrality of 13244.091), CSCE (KSZE, 10264.730) and foreign policy (AUSZENPOLITIK, 9404.719). However, the network graph reveals another story (cf. Figure 3), visualizing foreign policy as the most central concept in the report network. Since the graph is produced with parameters pushing hubs at the periphery and putting authorities more central (see also Table 1), communist party seems to be a hub instead. A closer analysis of the report data supports this interpretation: the keyword communist party co-occurs 77 times in 8 reports, whereas CSCE co-occurs 138 times in 12 reports and is thus a more central keyword than communist party. The keyword foreign policy has

\textsuperscript{31} Stuart & Botella, 2009, 15
\textsuperscript{32} Stuart & Botella, 2009, 15; Feicheng & Yating, 2014, 234
\textsuperscript{33} Hsu & Kao, 2013. See also Prell, 2012, 103-104
\textsuperscript{34} See also Paranyushkin, 2011, 13-14; Feicheng & Yating, 2014, 235
Table 2: Dynamics of the top-5 keywords and keyword co-occurrences (1975-1989).

<table>
<thead>
<tr>
<th>Keyword co-occurrence</th>
<th>N</th>
<th>Keyword co-occurrence</th>
<th>N</th>
<th>Keyword co-occurrence</th>
<th>N</th>
<th>Keyword co-occurrence</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (1975-1989)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSCE ↔ East-West</td>
<td>7</td>
<td>Domestic policy ↔ Northern Europe</td>
<td>5</td>
<td>CSCE ↔ EC</td>
<td>2</td>
<td>East-West relations ↔ Finland</td>
<td>3</td>
</tr>
<tr>
<td>CSCE ↔ Political-ideological diversion</td>
<td>6</td>
<td>Domestic policy ↔ Finland</td>
<td>5</td>
<td>CSCE ↔ Finland</td>
<td>2</td>
<td>East-West relations ↔ GDR</td>
<td>3</td>
</tr>
<tr>
<td>CSCE ↔ Rival actions</td>
<td>6</td>
<td>CSCE ↔ FRG</td>
<td>4</td>
<td>CSCE ↔ Northern Europe</td>
<td>2</td>
<td>Raw materials ↔ FRG</td>
<td>2</td>
</tr>
<tr>
<td>CSCE ↔ EC</td>
<td>5</td>
<td>CSCE ↔ Northern Europe</td>
<td>4</td>
<td>CSCE ↔ FRG</td>
<td>2</td>
<td>Raw materials ↔ Finland</td>
<td>2</td>
</tr>
<tr>
<td>CSCE ↔ Co-operation</td>
<td>5</td>
<td>CSCE ↔ Central Europe</td>
<td>4</td>
<td>CSCE ↔ East-West relations</td>
<td>2</td>
<td>Raw materials ↔ Switzerland</td>
<td>2</td>
</tr>
</tbody>
</table>

the least co-occurrences (74), but appears in 20 reports. In other words, the keyword foreign policy has fewer co-occurrences than the other central concepts, but it links more reports to each other. Further, the notion made by Feicheng & Yating (2014, 235) holds true also in our analysis: the keywords having the highest betweenness centrality have low closeness centrality, which measures a node’s independence in form of its closeness to other actors and vice versa. For example, the keyword foreign policy has the 8th lowest closeness centrality.

The previous analysis has focused on the overall network topology and represented the report corpus as a static network. However, historical research is mainly interested in changes in time, developments in a certain period of time and the dynamics behind or resulting in or from historical processes. In the recent years, network dynamics has gained in importance and new approaches and methods either to model change in network structures or test explanations for observed change have been constructed. The underlying assumption is that we can capture the dynamics of the report network by slicing the material in snapshots consisting of reports from five-year periods. The network data used in this analysis consists of both the keywords and country references of each report, and thus, allows the tracking down not only of the thematic dynamics, but also possible changes in geographical focus.

To start with the basic network characteristics, the network becomes slightly denser over time: 0.016 (1975-1979), 0.028 (1980-1984) and 0.031 (1985-1989). At the same time, however, the network size decreases first from 296 nodes and 696 edges (1975-1979) to 154/336 (1980-1984) and finally to 112/192 (1985-1989). At the same time, the average path length (2.954 / 2.911 / 3.161) and clustering coefficient (0.102 / 0.093 / 0.057) remain relatively stable. The last period (1985-1989) is quite interesting, since the number of reports is bigger than in 1980-1984, but at the same time the co-occurrence network is smaller and less dense, implying changes in the set of keywords used for summarizing the report contents. A closer analysis of the set of keywords reveals, that the average number of keywords per document changes from 13 (1975-1979) to 12 (1980-1984) and to 6 (1985-1989). At the same time, the total number of keywords used drops from 217 to 99 and then to 83.

These changes in the set of keywords clearly affect the content structure as well. Considering, first, the thematic pathways build around the core keywords, the most remarkable shift occurs in 1985 (Table 2). Until 1985, keywords related to CSCE dominate the co-occurrence network. This content structure well correlates with the historical fact that from the second half of the 1970s onwards the CSCE’s political consequences for the Communist camp began to dominate the GDR’s

35 For a good summary, see Scott, 2013, 139-145
political agenda. The party leadership in the GDR was increasingly worried about the destabilizing impact of the CSCE on its power and, consequently, instructed the HV A to continuous evaluation of the situation in Europe. Also in the early 1980s, CSCE remained central in the report corpus. However, during this period the CSCE was more and more embedded in the wider context of Western European integration, East-West divide and emerging contacts between socialist and capitalist countries. In 1984, the official visit of the Secretary General Erich Honecker in Finland dominated the reporting of the HV A and most of the reports evaluated both Finnish and West-German reactions before and after Honecker’s visit. The most significant change in the content structure occurs from 1985 onwards. First of all, the reports became more focused on Finland and Finland’s western relations. But the changes also imply that economic issues and questions related to energy production gained in importance. These changes also correlate well with historical developments: in the late 1980s bottlenecks in the GDR’s energy sector worsened dramatically as the Soviet Union cut its oil exports and due to the increase in international oil price. Together with general lack of raw materials, the question of finding alternative concepts and technologies for energy production and consumption - energy-saving included - gained a high priority also in the HV A’s activities. The party leadership increasingly expected the HV A to gather know-how and scientific-technical knowledge, an expectation the HV A sought to fulfill by intensifying its scientific-technical intelligence.

These changes in the content structure of the report corpus are supported by structural changes over time in the set of keywords used for summarizing the reports, as evidenced by the over 60 percent decrease in the total number of keywords. The content also changed radically. The set of keywords used between 1980 and 1984 contained 63 percent of the keywords used between 1975 and 1979. The set of keywords used from 1985 onwards contained 45 percent of the keywords used from 1975-1979, and only 39 percent of those used in the first half of the 1980s. However, the core keywords - CSCE, foreign policy, domestic policy - remain steady over time and can thus be regarded as the most important thematic path in the report corpus.

The keyword co-occurrence network visualizations for the most part support the results from the content structure analysis (Figure 4). There are, however, some observable differences between centrality metrics and positions in the graph worthy of discussion. In the graph showing the keyword co-occurrence network of 1975-1979, the keywords communist party and evaluation (label EINSCHAETZUNG) ranked as first and third according to betweenness centrality metrics and are pushed as hubs to the graph periphery (see Figure 4, sub-figure (a)). The graphs for 1980-1984 and 1985-1989 demonstrate a rather scattered co-occurrence structure, though they confirm the importance of keywords with the highest betweenness centrality metrics (see Figure 4, sub-figure (b) and (c)). This seems to be quite logical, since several keywords are included in approximately the same amount of reports, thus making it difficult for the layout algorithm to find clear authorities.

What is actually the story the network graphs visualizing the dynamics of the report corpus are telling about the historical period of 1975-1989? The changes detected in the thematic structure of the reports can be read as a map of challenges the GDR was facing. In the second half of the 1970s, the CSCE clearly dominated the political agenda of the GDR. From the early 1980s onwards, economic issues and questions related to international politics became more significant. The developments in the second half of the 1980s clearly illustrate how East German foreign intelligence became more and more harnessed for finding solutions to domestic problems of the GDR, mostly in the technological domain. Thus, the reporting of the HV A revolved to a greater extent around economic and energy-related issues, but also around West European integration. However, taking into account the limited number of reports in each period, the network graphs seem to suggest a shift away from thematically focused reports in the direction of general assessments of the current situation dealing with a wide range of questions. The visualizations also show that in regard to Nordic affairs the HV A was increasingly engaged in gathering information and compiling reports and analyses on economic problems and their solutions. Since these problems are known to have played an important role in the demise of communism, the HV A seems to have recognized the danger, yet not been capable of producing any usable solutions.

37 See Gieseke, 2001, 210-214
38 E.g. Wallander, 2003; Gaddis, 2005; Rafalzik, 2010
39 See also Müller-Enbergs, 2010, 111
Figure 4A: 1975-1979
Content Structure of Intelligence Reports

Figure 4B: 1980-1984
Content Structure of Intelligence Reports
Content Structure of Intelligence Reports

Figure 4D: Overall (1975-1989)
The issue of network clusters has already been touched upon above, as both the network visualizations and network metrics indicated the existence of different thematic clusters in the intelligence report corpus. In general, the growing availability of digitalized historical materials and on-line resources, data classification and categorization (including community, clique or sub-network detection) as methods to focus the research on critical points and structures have gained in importance. However, despite extensive studies, community detection has remained a core problem and theoretical challenge in network analysis. This article will utilize the community detection mechanism based on modularity also known as "Louvain method", which relies on the assumption that nodes being more densely connected to each other than with the rest of the network construct a cluster or community. Since Gephi has a built-in support for this method, its application is quite straightforward.

Community network metric imply the existence of three main content clusters. The largest cluster is comprised of 21.65 % of the total keywords and is built around the keywords CSCE, foreign policy and social democracy. The second and third content clusters include 20.96 % and 18.56 % of the total keywords, respectively. The second largest content cluster is characterized by the keywords domestic policy, raw materials and standpoint. As for the third content cluster, this cluster revolves around the keywords communist party, differences and peace movement. It is worth noting that two keywords - CSCE and communist party – which belong to different clusters have the highest degree, and thus, seem to have a stronger influence on the co-occurrence structure than other keywords. The two clusters built around these two keywords have different semantic structures, and

40 For excellent summaries of recent discussions, see Fortunato, 2010; Wu et al., 2013
41 See Blondel et al., 2008
42 https://sites.google.com/site/findcommunities/ [on-line. Last visited on 9th October, 2014].
43 Newman, 2006; Paranyushkin, 2011, 18. Cf. Du et al., 2008
consequently, differ also in content. The cluster revolving around CSCE seems to have its focus on Western influence and international politics whereas the cluster built around communist party seems to deal with threats and challenges within the Communist camp.

The graph above visualizes the modularized community structure of the keyword co-occurrence network in two different forms, where each community has its own color (see Figure 5). The sub-figure (a) is an ungrouped visualization, i.e. the keywords are colorized but not grouped. Already this visualization implies the existence of several thematic clusters in the report corpus. The content cluster structure becomes more clear in the second visualization based on grouped keywords (Figure 5, sub-figure (b)).

4. Concluding remarks

This article sought to exemplify how social network analysis could be applied to keyword co-occurrence networks in order to uncloak hidden content structures in a corpus of intelligence reports and to understand continuity and change within the corpus. The methodology was based on two assumptions. First, all kinds of texts can be presented as networks capable of capturing the most important content. Second, text networks can be analyzed as social networks of language in which relationships between concepts or words are analyzed in a similar manner as relationships between individuals in social networks. From this perspective, concept co-occurrence networks not only embody the most relevant information, but also allow comparisons in time, and thus, help to focus on continuity and change.

As the analysis presented in this paper shows, SNA offers powerful tools for identifying the most influential (central) keywords that function as junctions in the content structure, for detecting distinct communities present in the report corpus, and for comparing networks in order to examine change. Further, comparisons with prior historical knowledge illustrated the reliability of the results produced by the new methodology. The latter is especially important in historical network research. A critical contextual assessment- “Does this make sense in this context?” - can protect from overestimating but also from underestimating the results. One should keep in mind that historical network analysis change the way in which a historian looks at the material, but also affects what they see. Therefore, results challenging previous knowledge can be exactly that - new knowledge.

The results underline the role of visualizations in the process of producing and presenting new knowledge. On the one hand, graphs are an important tool for validating the results of the network metric analysis and for focusing the future research. Graphs can also be used for exploratory analysis; instead of using networks analysis to answer ready-formulated research questions or for hypothesis testing, one could also ”let the graphs speak” in order to explore new views on old data or to generate new hypotheses. On the other hand, visualizations also verified the important impact of visualization layouts for knowledge production. Layouts should be understood as different views on the same network, views produced by different algorithms and parameters. Unfortunately, in many publications, visualizations are rarely used for presenting the same network from different perspective, but just for graphically presenting the network. Since layouts alter our perception of the network, a proper understanding of the visualization algorithm is indispensable for a critical assessment of the visualized knowledge. A proper layout selection can result in new knowledge while an unsuitable layout can nullify the results.

Two crucial aspects might limit the use of computational methods and algorithms among historians (or humanists in general). The first one is related to education and research training; humanists are less familiar with computational science than their colleagues.
e.g. in the social sciences, and consequently, often more skeptical about the benefits of Digital Humanities. In this sense, this paper should be read as an encouragement for Digital Humanities inspiring colleagues to leave the “comfort zone” and to familiarize themselves with methods and tools available for computational research in the humanities. The second aspect is related to the question of suitable sources for network data. To be sure, data for small-scale networks can be entered manually. However, the analytical step towards network analysis of large-scale, complex networks can be taken only through digitalization and computational processing of historical sources, which is technically demanding, costly and time-consuming. Hopefully, this article has shown that Digital Humanities may be the multi-disciplinary window to a new analytical landscape, to a new understanding of historical events - not as a substitute for, but as a complementary to existing research techniques and methodologies.

References


Organisation Theory, 16(1), 89–111.


Rafalzik, Sascha. (2010). Wirtschaftsspionage der DDR. Münster: LIT.


American Foreign Policy Interests, 30(4), 196–210.


