Christina Niederer / Alexander Rind / Wolfgang Aigner /Julian Ausserhofer / Robert Gutounig / Michael Sedlmaier

Visual Exploration of Media Transparency for Data Journalists: Problem Characterization and Abstraction

109 - Data Science: Erfassung, Modellierung, Analyse und Visualisierung von Daten

Abstract

Today, journalists increasingly deal with complex, large, and heterogeneous datasets and, thus, face challenges in integration, wrangling, analysis, and reporting these data. Besides, the lack of money, time, and skills influence their journalistic work. Information visualization and visual analytics offer possibilities to support data journalists. This paper contributes to an overview of a possible characterization and abstraction of certain aspects of data-driven journalism in Austria. A case study was conducted based on the dataset of media transparency in Austria. We conducted four semi-structured interviews with Austrian data journalists, as well as an exploratory data analysis of the media transparency dataset. To categorize our findings we used Munzner´s analytical framework and the Data-User-Task Design Triangle by Miksch and Aigner.

Keywords:

Data-driven journalism, interactive data exploration, information visualization, user-centered design, visual analytics

1. Introduction

The independence of media companies from governmental interference is regarded as a cornerstone of democracy. Thus, Austrian governmental organizations are legally required to report the money flow for advertisement and media sponsoring, which are collectively published as open government data on media transparency (RTR 2015). This dataset represents a valuable resource for journalists. For example, they might investigate connections between advertisement by public bodies and state-owned enterprises and the presence of politicians in a newspaper.

At least since 2010, these practices have been publicly discussed under the term "data-driven journalism" (Lorenz 2010) (Gray et al. 2012). Other expressions, such as computational journalism, computer-assisted reporting and database journalism are also popular. Data-driven journalism encompasses both, the computational exploration of data, and the display of interactive visualizations in the news (Loosen et al. 2015). While the exploration of smaller datasets has already become a routine for many journalists, the investigation of more complex and bigger datasets is still a major challenge. The specific work done in the field of data-driven journalism is often characterized by comparing “values in order to show differences and similarities between different objects of study (e.g., people of different gender, neighbourhoods)” (Loosen et al. 2015). Although heavily needed for reporting, methods such as automated analysis, the integration of heterogeneous datasets and complex data wrangling are almost never employed due to lack of money, time and – above all – skills (Ausserhofer et al. 2015).

Information visualization and visual analytics (VA) (Keim et al. 2010: p. 7) offer solutions that “combin[e] automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets." However, the design space for possible visual representation, interaction, and automated analysis techniques is huge (Munzner 2015). Thus, a solution suitable for data journalists depends on a sound characterization of their work practices, analysis problem and requirements (Sedlmair et al. 2012).

Furthermore, abstraction is needed to bridge the needs between the language of the domain of journalism and VA techniques. In this paper, we contribute to such a better characterization and abstraction of data-driven journalism in Austria with a case study of Austrian media transparency.

2. Related Work

There are a number of studies that investigated the practices and conditions of data-journalistic news work. Much is known about the situation in major newsrooms - especially in the USA, Scandinavia, and the UK (Ausserhofer et al. 2015). Visualization techniques respectively tools, supporting data journalists in their work are less explored. Brehmer et al. (2014) presented a visualization design study of *Overview,* a tool for the systematic analysis of large document collections for journalists. They also conducted a task and data abstraction for visualization design methodology for the domain of journalism. The media transparency dataset itself has been visualized for several news websites (e.g., derStandard.at, Paroli Magazin) but these infographics show a small predefined subset of the data with minimal or no interactivity and are thus not suitable for exploration. FH Joanneum (2013) published an interactive online tool for interested citizens that can also be useful for data journalists, but it does not consider changes in money flow over time. Methods for visualization of data similar to the media transparency dataset – dynamic, weighted, and directed graphs – have been surveyed by Niederer et al. (2015) with a particular focus on data-driven journalism. The wider field of visualizing dynamic graphs is covered by Beck et al. (In Press).

Besides the data and task abstraction of Brehmer et al. (2014) with *Overview*, no research project is known that is concerned with the characterization and abstraction of problems and needs of data journalists regarding the visual exploration of dynamic, directed, weighted, bipartite networks data and with an exploration of the media transparency dataset in particular.

3. Method

For our problem characterization, we conducted four semi-structured interviews (Lazar et al. 2010) with four data journalists and performed an exploratory data analysis (Tukey 1977) of the media transparency dataset using Excel and Tableau. We used categorizations by Munzner (2015) for the identified tasks during the interviews. The interviews took approximately one hour and were conducted in person and via video call. The main aim of these interviews was to answer the following questions based on the Data-User-Task Design Triangle of Miksch and Aigner (2014): What kind of data are the users working with? (data) Who are the users of the VA solution(s)? (users) What are the (general) tasks of the users? (tasks)

The surveyed journalists have a more than 10-year experience in journalism. Beside their journalistic activities, three of the subjects teach journalism or data journalism at different universities of applied sciences in Austria. All four interviewees know the media transparency dataset but only one has actively worked with the dataset towards creating a news article.

4. Results

We summarize the findings structured around the three questions asked in the interviews. *Data:* The media transparency dataset can be modeled as a dynamic bipartite network of legal entities (e.g., a municipality) and media (e.g., a newspaper) in combination with quantitative flows (a dynamic, weighted, directed graph) (Beck et al. 2015). Figure 1 represents the dynamic bipartite network topology. The vertices fall into two sets - legal entities on the left and media institutions on the right. The edges are weighted and directed and link legal entity to one or more media vertices. There is no relationship within the vertices of the two sets.

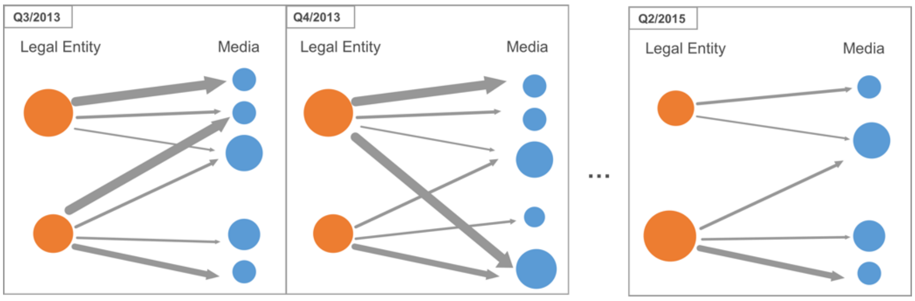

Figure 1: Data structure of the media transparency dataset, showing changes over time

Currently, data for 8 quarters (Q3/2013–Q2/2015) are available comprising 19,456 quarterly money flows (in total, without missing values). 916 distinct legal entities and 2,736 distinct media institutions occur in this period. The amount of money is available as a numeric value with a value range from € 5,000 to € 297,300,000 (average: € 98,379.59, median: € 10,931.92). Flows below € 5,000 are not registered and reported as missing values. Additionally, the legal background of the money flow is present as categorical attribute with 3 values (§2= advertising assignment and media cooperation, §4= funding to media owners, and §31= Announcement of the ORF delivered program fee) (RTR 2015).

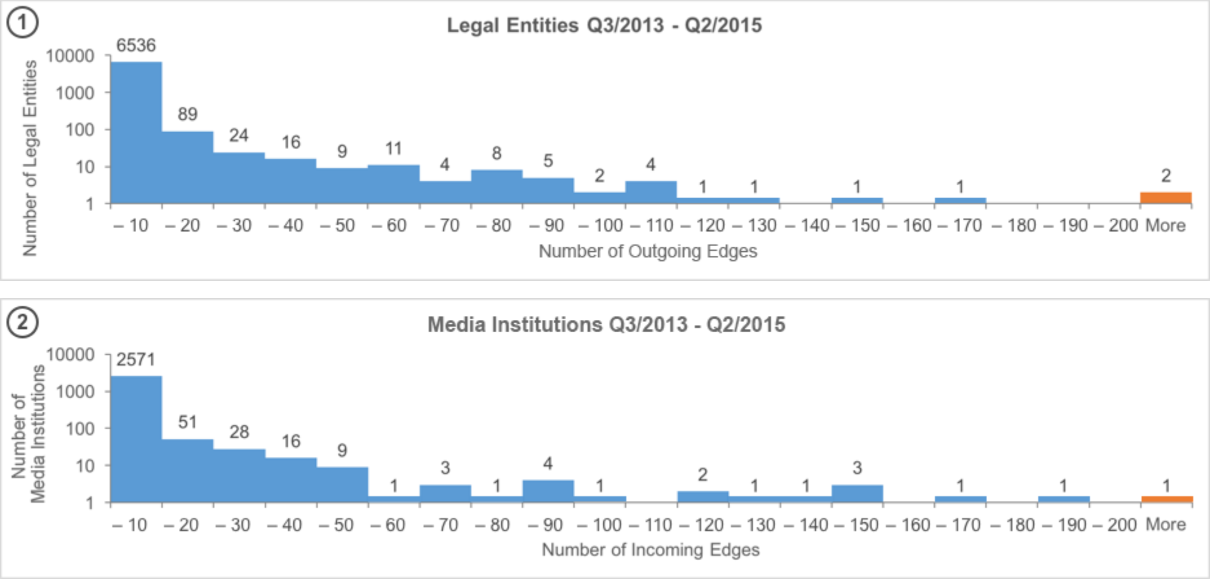

Figure 2: Histogram showing the number of outgoing and incoming edges (for 8 quarters) (bins of 10 and logarithmic scale on the y-axis) for (1) legal entities and (2) media institutions

Legal entities have on average 9 outgoing edges per quarter. Figure 2 (1) shows a histogram of legal entities by their outgoing edges over eight quarters. It is interesting to note that legal entities having only a single outgoing edge (i.e. one money flow to one media in one quarter) are most frequent with a count of 5,800. 736 legal entities have between two and 10 outgoing edges. Three legal entities have more than 150 edges to media: “Österreichische Werbung” (383), “Stadt Wien” (291), and “RTR-GmbH” (168). Media institutions have on average approximately 6 incoming edges per quarter. Figure 2 (2) represents the distribution of media by number of incoming edges, which is similar to outgoing edges of legal entities. 2,102 media institutions have only one incoming edge and 415 media have between two and 10 incoming edges. Three media institutions have more than 150 incoming edges: “Kronen Zeitung” (250), “Kurier” (170) and “ORF2” (185).

Users: According to our interviewees, data journalists in Austria have an average computing background as well as basic statistic knowledge. The exploration and research phase is organized depending on their personal preferences and skills. Thus, a workflow or specific user description could not be identified based on the conducted interviews. Due to the investigative character of their work, the interviewed journalists also have to deal with a wider range of data sources stemming from all kinds of entities in many other domains. Deadlines and lack of time (and resources) for extensive analysis are also limits for data-intensive news work (Karlsen and Stavelin, 2014). The interviews indicated that journalists are comfortable with combining different tools such as Excel, Google Spreadsheets for data analysis or other tools like Evernote for storing notes, found during the exploration or research phase. Further, the research documentation has been reported to be electronic as well as on paper. Regarding visualization techniques, standard business charts such as bar charts, line plots, or pie charts were mentioned as being known and used. The journalists expressed that they prefer easy to use interfaces and visualizations. It was reported, that the research phase on a news story is usually carried out alone. The next steps of designing infographics as well as drafting the final article is then done in teams of 2 to 3 people.

Tasks: Based on the what-why-how framework of Munzner (2015) the tasks identified in the interviews and data analysis are described and categorized. The primary task of the interviewed journalists is to analyze existing datasets to discover outliers and features. As a secondary task, journalists look up the data to verify an existing hypothesis and/or explore the datasets to find any interesting parameter. The interviewees explained that abnormality in the media transparency data relating to minimum or maximum flows of money or patterns recognizable over a period of time could be a starting point for the exploration phase. All four subjects mentioned that changes in the data, for example the varying number of ingoing or outgoing edges based on one special legal entity or media institution are of interest. Different tasks during browsing are comparing and summarizing different datasets and values. Also different annotation and recording tasks are part of the exploration process. Interviewee 2 commented that adding connections manually based on the journalistic experience and knowledge could be a helpful feature. Also the subjects suggested a manual or automatic grouping functionality as a relevant feature. This was also suggested by the other interviewees. Analytical provenance, a documentation of the research path, and a feature to create screenshots/snapshots of the research status were indicated as useful.

5. Conclusion

The problem characterization and abstraction at hand constitutes a necessary step towards developing visualization and VA methods that suit the needs of data-driven journalism. The automatic or manual grouping or linking of data based on the journalistic background knowledge and experience or integration of further data sets can be defined as one important need. Also the documentation of the research path of the exploration phase can be ranked as a major need of data journalists. A second finding is that journalists are interested in changes of the data over time and in finding abnormalities such as large flow of money or recurring pattern. Because of limited statistical and programming skills, the developed tools have to be very easy-to-use and self-explaining. The requirements such as linking heterogeneous data, work under time pressure, or preference for common well-known visualization techniques have parallels in many other domains. Further studies should be carried out to explore if a tool for data journalists can be generalized for other domain experts with the same type of data basis (dynamic, directed weighted bipartite network).

Acknowledgements: This work was supported by the Austrian Ministry for Transport, Innovation and Technology (BMVIT) under the ICT of the future program via the VALiD project no. 845598 and by the Austrian Science Fund (FWF) via the KAVA-Time project no. P25489. Many thanks to our interviewees for their inputs and to Markus Wagner for his feedback.

References:

Aigner, W., Miksch, S., Schumann, H., Tominski, C., 2011. Visualization of Time-Oriented Data, Human-Computer Interaction Series. Springer, London.

Ausserhofer, J., Gutounig, R., Oppermann, M., 2015. News production workflows in data-driven, algorithmic journalism: A systematic literature review. [WWW Document]. URL http://www.validproject.at/wp-content/uploads/2015/11/151031_AusserhoferGutounigOppermann-Dubrovnik_Dubrovnik.pdf

Beck, F., Burch, M., Diehl, S., Weiskopf, D., In Press. A Taxonomy and Survey of Dynamic Graph Visualization. Comput. Graph. Forum.

Brehmer, M., Ingram, S., Stray, J., Munzner, T., 2014. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. IEEE Trans. Vis. Comput. Graph. 20, 2271–2280. doi:10.1109/TVCG.2014.2346431

FH JOANNEUM Gesellschaft mbH, 2013. Medientransparenz Austria [WWW Document]. Medien. Austria. URL http://www.medien-transparenz.at/ (accessed 12.14.15).

Gray, J., Chambers, L., Bounegru, L., 2012. The Data Journalism Handbook, 1st ed. O'Reilly and Associates.

Karlsen, J., Stavelin, E., 2014. Computational Journalism in Norwegian Newsrooms. Journal. Pract. 8, 34–48. doi:10.1080/17512786.2013.813190

Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (Eds.), 2010. Mastering The Information Age – Solving Problems with Visual Analytics. Eurographics Association, Goslar.

Lazar, D.J., Feng, D.J.H., Hochheiser, D.H., 2010. Research Methods in Human-Computer Interaction. John Wiley & Sons.

Loosen, W., Reimer, J., Schmidt, F., 2015. When Data Become News: A Content Analysis of Data Journalism Pieces. Presented at the The Future of Journalism 2015 Conference, Cardiff.

Lorenz, M., 2010. Status and outlook for data-driven journalism, in: Data-Driven Journalism: What Is There to Learn? European Journalism Centre, Amsterdam, pp. 8–17.

Miksch, S., Aigner, W., 2014. A Matter of Time: Applying a Data-Users-Tasks Design Triangle to Visual Analytics of Time-Oriented Data. Comput. Graph. 38, 286–290. doi:10.1016/j.cag.2013.11.002

Munzner, T., 2015. Visualization Analysis and Design. CRC Press.

Niederer, C., Aigner, W., Rind, A., 2015. Survey on Visualizing Dynamic, Weighted, and Directed Graphs in the Context of Data-Driven Journalism, in: Schulz, H.-J., Urban, B., Lukas, U.F. von (Eds.), Proceedings of the International Summer School on Visual Computing 2015. Fraunhofer Verlag, Rostock, pp. 49–58.

RTR [WWW Document], 2015. URL https://www.rtr.at/de/m/Medientransparenz (accessed 9.16.15).

Sedlmair, M., Meyer, M., Munzner, T., 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. IEEE Trans. Vis. Comput. Graph. 18, 2431–2440. doi:10.1109/TVCG.2012.213

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA.