EuroVis Workshop on Visual Analytics (2017)
M. Sedlmair and C. Tominski (Editors)

Guidance for Multi-Type Entity Graphs from Text Collections

M. Müller [1], K. Ballweg[1], T. von Landesberger[1], S. Yimam[2], U. Fahrer[2], Ch. Biemann[2], M. Rosenbach[3], M. Regneri[3], H. Ulrich[3]

[1]Interactive Graphics Systems Group, Computer Science, TU Darmstadt, Germany
[2]Language Technology Group, Department of Informatics, Universität Hamburg, Germany
[3] Spiegel-Verlag, Germany

Abstract

The visual exploration of graphs encoding relationships between entities of multiple types (e.g., persons, locations,...) supports journalists in finding newsworthy information in large text collections. Journalists may have interest in certain entity types or their relations such as locations or person-person relations. This interest may change during the exploration process. The exploration of such large graphs is often supported by guidance using a degree-of-interest (DOI) function. Although many DOIs exist, they do not differentiate entity types, rely on additional data, or require complex settings overburding the journalists.
We present a novel DOI for graphs with multiple types of entities. We show the interesting subgraph around the focal node and offer information about possible further steps. The user can interactively set her interest in entity types and entity relations. We apply our approach to a graph extracted from WikiLeaks PlusD Cablegate documents and report on journalists' feedback.

Categories and Subject Descriptors (according to ACM CCS): Interaction, Visual Analytics, Guidance, Graphs, Digital Humanities

1. Introduction

Journalists wish to find interesting information in large document collections such as Kissinger Cables to derive newsworthy stories. As reading all the documents is very time consuming, they explore so-called entity graphs extracted from the documents. The graphs contain entities (e.g., persons) and the relations between these entities (i.e., co-occurence in documents). These graphs may contain various types of entities such as persons, locations, organizations. They are called multi-type entity graphs.

Our requirement analysis showed that journalists at the newspaper SPIEGEL often analyze unknown document collections following an open exploration process [BZWD*16]. The journalists may focus on certain entity types or entity connections (e.g., persons, or person-organization relation). Their focus changes during this process depending on the discovered information. For example, they may be first interested in a general overview of entities connected to Angela Merkel. Then, they wish to analyze only on person-person relations (such as Merkel-Obama) and then person-location relations (such as Merkel-USA) .

Such exploratory process is currently very time-consuming due to the large datasets, i.e., graphs having millions of nodes. For example, the graph extracted from WikiLeaks PlusD (Cablegate) documents [Wik] has 1.4 million entities of four types (persons, locations, organizations and miscellaneous). Entities are connected by 163 million edges representing entity co-occurrence in documents.

The exploration of large entity graphs is often supported by guidance. Based on a degree-of-interest (DOI) function, the guidance shows the user an interesting subgraph of the whole graph and offers suggestions for further exploration steps. Although many approaches exist (see Sec. 2), they do not distinguish entity types [vHP09, KvLB14], require additional information [MSDK12] or rely on complex settings overburdening the journalists [AHSS14].

We present a novel algorithm for guidance in multi-type entity graphs and combine it with an interactive user interface (see Fig. 1 & 2). We show the interesting subgraph around the user's focal node and inform the user about possible exploration steps. We also offer simple but effective ways of setting user interest in entity types or in types of entity relations (e.g., person-location). Our approach extends the *new/s/leak* system for [BZWD*16, YUvL*16], which offers interactive features for close and distant reading.

We apply our approach to a graph extracted from a WikiLeaks PlusD texts. We show an example use case. We also gained feedback from journalists at SPIEGEL. It confirmed the value-added of the entity-type preference in the guidance as well as the easy usage of the visual interface for the journalistic research.

2. Related Work

Guidance offers an environment, which supports the users in performing their tasks to progress in their analytical goal [CGM*17]. Various means of guidance are described in [CGM*17]. Our work focuses on directing journalists towards finding relevant data subsets in an entity graph. The directing shows relevant options for reaching the goal. The guidance input are both the entity graph data and user's preferences. The guidance outputs are relevant data

DOI: 10.2312/eurova.20171111

subsets – subgraphs of the original graph, which are shown on the screen. Such guidance often relies on a so-called a degree-of-interest function (DOI). We review relevant works this area.

Based on early works by Furnas [Fur86, Fur06] and Card [CN02], van Ham and Perer [vHP09] presented a guidance system for the exploration of jurisdictional documents. They propose a node-based DOI, which contains apriori interest (API), distance to the focus node (D) and user interest (UI). They combine it with a visualization that offers visual clues for orientation. This approach has only one type of nodes. We extend it for multi-type nodes.

May et al. [MSDK12] use signposts to offer user orientation in the graph extracted from medical papers. Their guidance requires graph clustering, which is not available in our case. Calculation of clusters, see [GKN05, vHvW04], may produce clusters without specific meaning [CGM*17]. Thus, we refrain from it.

Abello et al. [AHSS14] proposed a modular DOI for nodes and edges. The DOI parameters are set in a user interface. It shows the distribution of the many DOI components and combines them. This flexibility requires a high level of user expertise. However, journalists have low visual literacy and need easy to use tools [BZWD*16].

Our previous work presented a tool for exploring an entity graph extracted from news articles [KvLB14]. It proposed DOI based on edges, not on nodes like [vHP09], because the latter resulted in the same subgraph irrespective of the focal node. However, this DOI disregarded various types of entities. Now, we extend this work in two ways. First, we include the user preference for certain entity types and their relations. Second, our approach can deal also with dense networks, where the previous approach resulted in "star-shaped" subgraphs around the focal node (see Fig. 2 a and b).

3. Approach

Our guidance approach has two parts: 1) algorithmic calculation of the degree-of-interest (DOI) subgraph around the focus node and 2) graph visualization enhanced with hints for further graph exploration. These two parts are generally found in guidance approaches. The novelty of our approach lies in 1) a novel DOI algorithm for multi-type entity graphs and 2) the enhancement of graph visualization interface with entity-type preference setting and specific hints for the exploration by entity type. These two parts extend the existing system *new/s/leak* [BZWD*16, YUvL*16] for the exploration of large document collections (see Fig. 1).

3.1. Degree of Interest Function

Our algorithm uses a degree-of-interest (DOI) function that has three main parts [vHP09]: apriori interest *API*, distance to the focal node D and user interest (UI), see Eq. 1. Our UI has two components: 1) user's preference for certain entity types and for connection types UI_{type} and 2) browsing history UI_{his}. Building upon our work [KvLB14], we define the DOI over edges, not nodes. The DOI for an edge between nodes x and y given a focal node z and user preference for entity types w is in Eq. 1:

$$DOI(\{x,y\}) = API(\{x,y\}) + D(\{x,y\}|z) + UI_{type}(\{x,y\}|w) + UI_{his}(\{x,y\}|w) \quad (1)$$

We use a single-focus DOI, as we assume that journalists forage from one interesting story to the next. They wish to find and explore

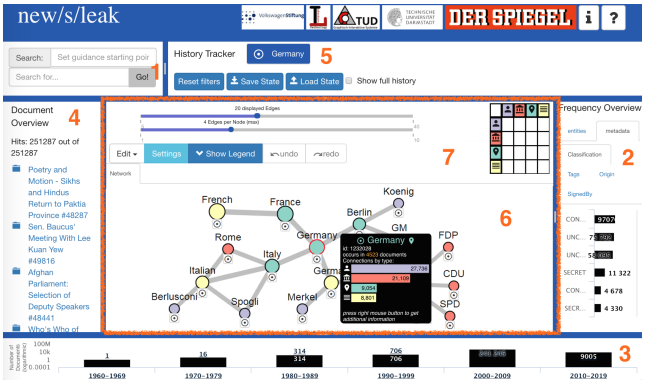
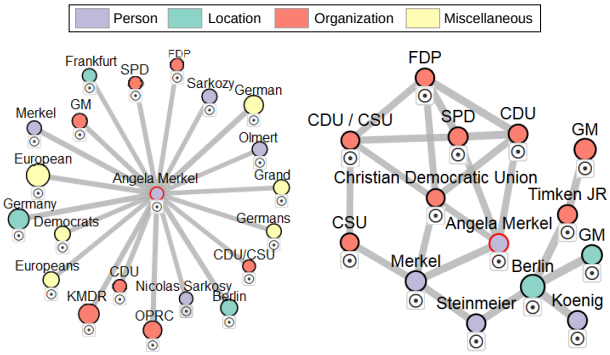


Figure 1: *New/s/leak system with guidance. Features: search (1), entity frequencies (2), timeline (3), document list (4), history (5), graph view with context menus (6) and guidance settings (7).*

interesting entities (focus nodes) and their relations for which they then study the related documents in detail.

Apriori Interest (*API*): Our API function stayed the same as in [KvLB14]. The edge apriori interest is given by the normalized Pointwise Mutual Information (*npmi*), which reflects the co-occurrence of the entities x, y in the documents *docOcc* (see Eq. 2).

$$npmi(x,y) = \frac{docOcc(x,y)}{docOcc(x) \times docOcc(y)} / (-ln(docOcc(x,y))) \quad (2)$$



(a) Standard DOI by [KvLB14] (b) Our approach

Figure 2: *The DOI subgraph for the focus node "Angela Merkel". a) The state-of the-art approach by [KvLB14] results in a star-shaped subgraph hiding interesting structures and entities. b) Our approach shows a better structured subgraph.*

Distance to the Focal Node (D): The distance to the focal node varies in DOI functions [AHSS14, vHP09, KvLB14, HZ07, MSDK12, CLWM11]. We are inspired by the most relevant works: [KvLB14, vHP09]. They use the exponential function of the unweighted shortest graph distance *gdist* with $b = 0.5$ (see Eq. 3).

$$D(\{x,y\}|z) = -(1 - b^{gdist(\{x,y\}|z)}) \times API(\{x,y\}) \quad (3)$$

We needed to extend this approach as, in our case, it often resulted in a star-shaped graph (see Fig. 2a). The reason is that our entity graphs is denser then in previous cases. Our average degree is 120, much larger then 11 in [KvLB14]. Our extension limits the number of edges per DOI subgraph. Only the *dn* most interesting

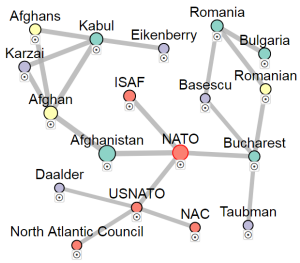
edges for a node are shown. We get a more structured DOI subgraph of a user-defined size s and degree dn (see Fig. 2b).

User Interest for Entity Types (UI_{type}): This is the novel part of our DOI. UI_{type} incorporates the user's preference w for certain entity types (e.g., $type$=persons) or entity relations (e.g., $type$=person-location), see also Eq. 4. We differentiate four preference levels: from no interest (disregard, $w = -\infty$), via normal interest ($w = 0$), high interest ($w = 0.05$) up to very high interest ($w = 0.25$). We empirically determined the values leading to meaningful results for our use case (see Fig. 2 b,c, and d). Note that for other graphs, the values may need to be adjusted. The user preference is defined for all entity types and relations in the user interface (see Section 3.2).

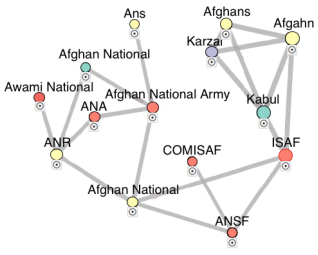
$$UI_{typ}(\{x,y\},w) = API(\{x,y\}) \times w(typ,\{x,y\}) \quad (4)$$

User Interest According to the Browsing History (UI_{his}) This DOI component ensures that the selection of a new focus node still preserves the context of the older browsing steps. Most interesting nodes and edges from the older browsing steps are included in the new subgraph. Figure 3 shows an example of graph browsing from NATO to ISAF. Without UI_{his}, most nodes are new, while the usage of UI_{his} ensures that many relevant nodes remain. Our UI_{his} includes decayed values of the older DOI in the more recent DOI calculation (see Eq. 5). The decay factor $0 < \beta < 1$ ensures that previous steps get less importance than new steps. After experimenting with various values, we use $\beta = 0.08$.

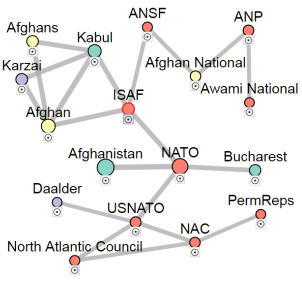
$$UI_{his}(x,y|z,w)^{new} = \beta \times DOI^{old}(x,y|z^{old},w^{old}) \quad (5)$$



(a) Old focus: NATO



(b) New focus: ISAF, No history



(c) New focus: ISAF, With history

Figure 3: *The need for user history in DOI function.*

3.2. Visual Interface

We extend the visual interface of the *new/s/leak* system for visual exploration of entities and their relations extracted from text documents (see Fig. 1). The system allows for data exploration using the entity search (1), filtering for interesting entities by their frequency (2), or document time (3), or exploring the subgraph around a user-specified focus node with a red border in the graph view (6). The graph is layouted using a force-directed algorithm. The user can select documents for close reading in the document list (4). All user actions are tracked and shown in the history view (5).

The graph view uses the DOI-based subgraph selection and introduces new visual features for guidance: 1) interactive setting of user preferences for entity types and highlighting the preferred edges in the subgraph 2) features for guided graph exploration: setting the focus node and expanding graph with help of guidance hints – special node and edge context menus (see Fig. 1 (6), (7)).

Interactive Setting of User Preferences: The user can interactively set the preferences for subgraph calculation in the settings view (see Fig. 1 (7)). The preferences for entity types and types of entity relations are set in the preference matrix (see Fig. 4 (2a), (2b)). It shows the relations between entity types. Each click on the matrix cell changes the preference level for this relationship type circularly from low to high. Clicking on the entity type changes the preference for all relationships to this entity (i.e., whole row/column). These preference changes start the subgraph recalculation. The resulting preferred edges are highlighted with blue color (see Fig. 4). Moreover, the user can use sliders to set the subgraph size and the maximum number of adjacent edges in the DOI subgraph.

Guided Graph Exploration: The guidance shows the DOI subgraph around the user-defined focus node (highlighted with red border). The user can set the focus node in various ways: 1) by searching for entities, 2) by selecting an entity according to its frequency or 3) by selecting an entity in the visible subgraph (using the circular button below the node). Setting a new focus node recalculates the DOI and shows the new subgraph around the new focus node. The user can also expand the graph with respect to the adjacent edges to a node using the context menu.

We developed specialized context menus to guide the user (see Fig. 4 "Berlin"-, "Bosniaks - Murphy" context menu). The node context menu shows information about the node type, node frequency and its connections by entity type. It also shows information about the results of focusing or expanding the node. We preview the four most interesting entities that will be shown after the action. The edge context menu shows the edge frequency, edge type and allows the user to change the preference for this edge type. All exploration actions are tracked and shown in the history view on the top of the system window (see Fig. 1 (5)).

4. Evaluation

We present an example use case and the results of user feedback. Both are based on the entity graph extracted from about 250,000 WikiLeaks PlusD (Cablegate) documents [Wik] using the natural-language processing described in [YUvL*16]. The graph has 1,363,500 entities of four types: persons (592,690), locations (179,210), organizations (431,889) and miscellaneous (159,711). The entities are connected by 163,137,632 edges.

User Feedback: Six experts from the journalistic domain working at SPIEGEL evaluated our tool. The session started with a 15-minute demonstration of our tool. Then, the experts could use the

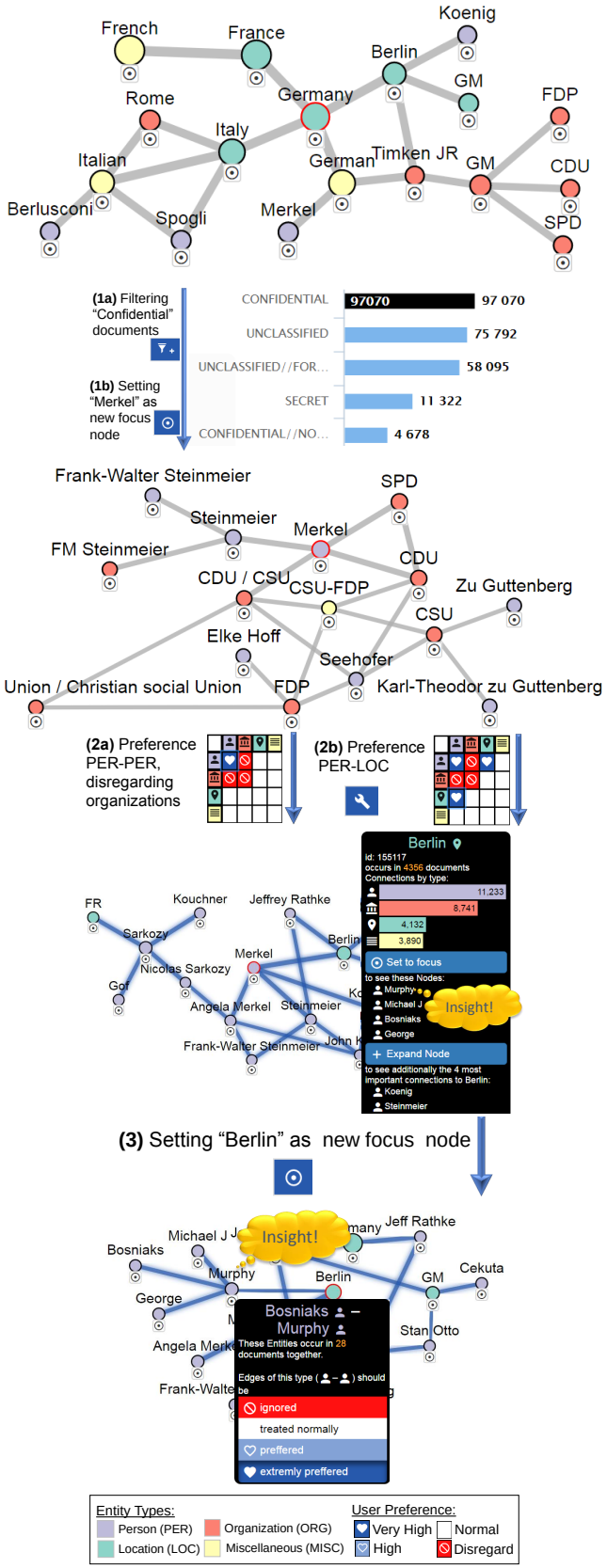


Figure 4: *Use case: political relations of Angela Merkel.*

tool in a one-hour hands-on session. At the end, they filled SUS forms [BKM08, Bro96] and gave free text feedback.

Already during the demonstration, the experts mentioned that the new DOI provides a value added. It is a novel feature not available in other journalistic tools. The forms showed that the advantage is especially relevant for browsing unknown data collections, as it supports the search for important information (5 experts). Moreover, it is a good extension to the standard *new/s/leak* document browsing features (2 experts). Two experts especially praised the value added of the exploration by the type of entity. One of them liked very much the matrix-based setting of the preferences and found the tooltip menu very intuitive for browsing interaction.
The SUS scores showed that the experts would like to use the system regularly, it is easy to use and integrated (all medians: 4, best: 5). The users do not need a lot of learning and do not require extra help to use the tool (both medians: 1, best: 1).

Example Use Case: The use case is shown in Figure 4 and in the video. A journalist is interested in exploring the confidential information about Angela Merkel. As expected, Angela Merkel is connected to many German politicians and parties. The journalist then prefers to see more person-person and person-location relations, while disregarding organizations. This reveals more politicians connected to Merkel such as N. Sarkozy. For further exploration, the journalists assesses the guidance in the context menus. She sees that setting Berlin to focus reveals connection to Mr. Murphy, who was a political consultant inter alia to Mitt Romney. Then, the updated subgraph shows interesting information about relations within Bosnian war. Thus, the journalist wishes to get more information by reading the related documents in new/s/leak tool.

5. Conclusions and Future Work

We presented a novel approach for the guided exploration of multi-type entity graphs, which are extracted from large document collections. Our approach is based on an extended DOI definition, which includes the user preferences for certain types of entities and their relations. The DOI determines a subgraph around a focal node visualized in a node link diagram. It also offers hints for further exploration (expansion or setting a different focus).

In the future, we will implement the user feedback and improve the stability of the graph layout during the exploration process. We could also incorporate more visual clues for further orientation such as those proposed in [vHP09, GST13]. We wish to extend the guidance also for the time-dependent graphs (e.g., changing document collections over time, or different time periods) building on [AHSS14, CSP*06, RPD09]. Our system uses single focus DOI being in line with user expectations. Nevertheless, in different use cases, multi-focal DOI could be advantageous, such as in [Osa01].

References

[AHSS14] ABELLO J., HADLAK S., SCHUMANN H., SCHULZ H.-J.: A modular degree-of-interest specification for the visual analysis of large dynamic networks. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2014), 337–350. 1, 2, 4

[BKM08] BANGOR A., KORTUM P. T., MILLER J. T.: An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction 24*, 6 (2008), 574–594. 4

[Bro96] BROOKE J.: SUS —A quick and dirty usability scale. *Usability evaluation in industry 189*, 194 (1996), 4–7. 4

[BZWD*16] BALLWEG K., ZOUHAR F., WILHELMI-DWORSKI P., FAHRER U., PANCHENKO A., YIMAM S., BIEMANN C., REGNERI M., ULRICH H.: new/s/leak – a tool for visual exploration of large text document collections in the journalistic domain. In *Visualization in Practice, VIS Workshop* (2016). URL: http://www.gris.tu-darmstadt.de/research/vissearch/publications//pdf/ballweg2016VIP.pdf. 1, 2

[CGM*17] CENEDA D., GSCHWANDTNER T., MAY T., MIKSCH S., SCHULZ H.-J., STREIT M., TOMINSKI C.: Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 111–120. 1, 2

[CLWM11] CRNOVRSANIN T., LIAO I., WU Y., MA K.-L.: Visual recommendations for network navigation. *Computer Graphics Forum 30*, 3 (2011), 1081–1090. 2

[CN02] CARD S. K., NATION D.: Degree-of-interest trees: A component of an attention-reactive user interface. In *Working Conference on Advanced Visual Interfaces* (2002), ACM, pp. 231–245. 2

[CSP*06] CARD S. K., SUH B., PENDLETON B. A., HEER J., BODNAR J. W.: Time tree: Exploring time changing hierarchies. In *Symposium On Visual Analytics Science And Technology* (2006), IEEE, pp. 3–10. 4

[Fur86] FURNAS G. W.: Generalized fisheye views. In *SIGCHI Conference on Human Factors in Computing Systems* (1986), vol. 17, ACM. 2

[Fur06] FURNAS G. W.: A fisheye follow-up: further reflections on focus+context. In *SIGCHI conference on Human Factors in Computing Systems* (2006), ACM, pp. 999–1008. 2

[GKN05] GANSNER E. R., KOREN Y., NORTH S. C.: Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 457–468. 2

[GST13] GLADISCH S., SCHUMANN H., TOMINSKI C.: Navigation Recommendations for Exploring Hierarchical Graphs. In *Advances in Visual Computing: International Symposium on Visual Computing* (2013), vol. 8034 of *Lecture Notes in Computer Science*, Springer, pp. 36–47. 4

[HZ07] HÜSKEN P., ZIEGLER J.: Degree-of-interest visualization for ontology exploration. In *IFIP Conference on Human-Computer Interaction* (2007), Springer, pp. 116–119. 2

[KvLB14] KOCHTCHI A., VON LANDESBERGER T., BIEMANN C.: Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles. *Computer Graphics Forum 33*, 3 (2014), 211–220. 1, 2

[MSDK12] MAY T., STEIGER M., DAVEY J., KOHLHAMMER J.: Using signposts for navigation in large graphs. *Computer Graphics Forum 31*, 3pt2 (2012), 985–994. 1, 2

[Osa01] OSAWA N.: A multiple-focus graph browsing technique using heat models and force-directed layout. In *International Conference on Information Visualisation* (2001), IEEE, pp. 277–283. 4

[RPD09] REITZ F., POHL M., DIEHL S.: Focused animation of dynamic compound graphs. In *International Conference on Information Visualisation* (2009), IEEE, pp. 679–684. 4

[vHP09] VAN HAM F., PERER A.: Search, show context, expand on demand: supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009). 1, 2, 4

[vHvW04] VAN HAM F., VAN WIJK J. J.: Interactive visualization of small world graphs. In *Symposium on Information Visualization* (2004), IEEE, pp. 199–206. 2

[Wik] WIKILEAKS: Public libary of US diplomacy. https://wikileaks.org/What-is-Wikileaks.html. 1, 3

[YUvL*16] YIMAM S. M., ULRICH H., VON LANDESBERGER T., ROSENBACH M., REGNERI M., PANCHENKO A., LEHMANN F., FAHRER U., BIEMANN C., BALLWEG K.: new/s/leak – information extraction and visualization for investigative data journalists. In *ACL System Demonstrations* (2016), Association for Computational Linguistics, pp. 163–168. URL: http://anthology.aclweb.org/P16-4028. 1, 2, 3