




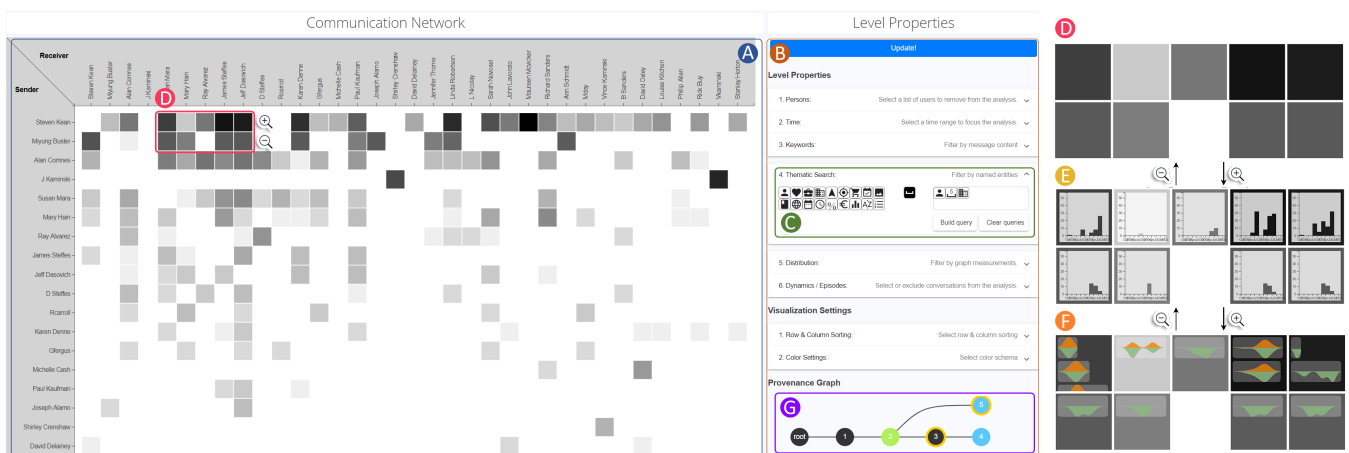


# CommAID: Visual Analytics for Communication Analysis through Interactive Dynamics Modeling

M. T. Fischer<sup>1</sup> , D. Seebacher<sup>1</sup> , R. Sevastjanova<sup>1</sup> , D. A. Keim<sup>1</sup> , and M. El-Assady<sup>1</sup> 

<sup>1</sup>Universität Konstanz, Germany



**Figure 1:** COMMAID, an integrated visual analytics technique to analyze communication networks through dynamics modeling, semantic pattern retrieval, and an interactive multi-level matrix-based visualization **A**. This visualization enables the inspection of individual communication at different analysis levels through semantic zooming, while the linked level property pane **B** allows for restricting the search space, using standard task filters, dynamics settings, and a thematic concept builder **C**. The in-line views show the visualizations provided by the individual analysis levels (different zoom steps) presenting volume information **D**, statistical distribution information **E** as well as communication episodes using conversational dynamics **F**. A provenance history graph **G** allows to keep track of the analysis steps and results. The technique allows to interactively explore communication activity both from a metadata (connectivity, closeness, time, statistics) as well as a content level (keywords, thematic concepts) simultaneously, reducing discontinuities.

## Abstract

Communication consists of both meta-information as well as content. Currently, the automated analysis of such data often focuses either on the network aspects via social network analysis or on the content, utilizing methods from text-mining. However, the first category of approaches does not leverage the rich content information, while the latter ignores the conversation environment and the temporal evolution, as evident in the meta-information. In contradiction to communication research, which stresses the importance of a holistic approach, both aspects are rarely applied simultaneously, and consequently, their combination has not yet received enough attention in automated analysis systems. In this work, we aim to address this challenge by discussing the difficulties and design decisions of such a path as well as contribute CommAID, a blueprint for a holistic strategy to communication analysis. It features an integrated visual analytics design to analyze communication networks through dynamics modeling, semantic pattern retrieval, and a user-adaptable and problem-specific machine learning-based retrieval system. An interactive multi-level matrix-based visualization facilitates a focused analysis of both network and content using inline visuals supporting cross-checks and reducing context switches. We evaluate our approach in both a case study and through formative evaluation with eight law enforcement experts using a real-world communication corpus. Results show that our solution surpasses existing techniques in terms of integration level and applicability. With this contribution, we aim to pave the path for a more holistic approach to communication analysis.

## CCS Concepts

• **Human-centered computing** → **Visual analytics**; • **Applied computing** → **Law, social and behavioral sciences**;

## 1. Introduction

The enormous growth in the use of electronic devices and systems in the past decades has led to an exponential increase in digital forms of communication. Simultaneously, the abundance of this digital communication [Sco09] and corresponding datasets has increased interest in how such communication can be analyzed in a wide variety of different domains, ranging from social sciences and digital humanities to engineering and business. For example, it has been studied how social and psychological features change with computer-mediated communication [GA08], how team performance can be assessed based on communication [FM08], how networks can be analyzed using text-mining [YP04], or how the evolution of dynamic communication networks can be visualized [Tri08]. This short list already shows a peculiar oddity when studying automated, digital communication analysis *systems*: most existing approaches focus on either the content of communication *or* on the network aspect—but not both. The first group of approaches usually leverages methods from natural language processing [MS99], while the latter uses techniques from the field of Social Network Analysis [Sco17].

Interestingly, this distinction between content and network is neither present in the seminal works on human communication research [Mor34, WBJ74], nor in modern works [Pea11] or current studies [FM08, Mes09], even if digital methods aid the manual analysis. Indeed, the analysis of network structure, communication patterns *as well as* content plays an integral part [SD60, WBJ74] of this research field. Abstractly, communication can be described as the exchange of meaning between entities, transporting information encoded into symbols [Pea11], reflecting the content's meaning and the network and metadata aspects through transport modalities. As such, analysis of the network/metadata or the content *alone* can sometimes provide a limited, incomplete, or even biased view on the communication, which is not ideal. Alternatively, employing independent approaches would introduce domain discontinues and complicates search tasks, placing an additional burden on the users.

The problem of how both the network and the content perspective can be combined has not yet received enough attention when considering communication analysis systems. This is especially relevant when such systems are used by non-communication experts, like in business intelligence applications or targeted criminal investigations, on which we based a case study (see Section 6.1). Typical tasks in these domains include searching for specific semantic content (e.g., negative product reviews, location names together with keywords), identifying groups (e.g., clusters or cliques), or a particular communication pattern (e.g., sequence, volume, timeline). These tasks can be addressed with separate solutions. But if, rather commonly, the search for specific semantic content needs to be restricted to a particular set of users (e.g., specific communication patterns, high centrality, or part of a specific clique), separate solutions struggle or even fail. Several other common tasks would benefit from such a combined search for cross matches and a fine-grained analysis of the communication network structure and context to detect such communication behavior and are therefore also not adequately supported so far. For example, detecting a terrorist attack on a soccer match by identifying the perpetrators increased chatter (hiding between the fans' increased chatter about the kick-off) about a parcel, which combines network, communication pattern, and content analysis.

In this work, we aim to address these shortcomings by discussing a possible technique as well as provide a framework for a holistic approach to interactive communication analysis. We do not aim to describe a turnkey system or replace existing solutions, but rather discuss the challenges and design decision in such a system, present an exemplary blueprint prototype on how such an integrated system could look like, and gather expert feedback on such a broader approach, to support further research and positively influence system development in this domain. For this, we extend upon two previous works, which we use as building blocks. The first work [SFS\*19] uses conversational dynamics to analyze communication patterns, covering the network analysis side. The second work [FAS\*20] describes a technique for hypergraph analysis, combining machine learning and a multi-level matrix-based visual interface. From this, we borrow and adapt parts of the visual interface technique: The conversational dynamics aspects are integrated in this framework as one analysis model, while the multi-level matrix-based interface design is repurposed for the main interface.

In this work, we present **COMMAID** (**Communication Analysis through Interactive Dynamics**), making the following contributions:

- A blueprint for a novel, interactive framework for a more holistic communication network analysis building upon individual models, providing a tight coupling between the network and the content analysis aspects.
- A description of two extendable models as example levels for network and content analysis, offering conversational dynamics [SFS\*19] and semantic concept detection, also including with several standard levels.
- A discussion on the challenges, design choices, and future work for a holistic communication approach.
- One case study and an assessment with eight law enforcement experts using real-world communication data describing an application in the law enforcement field.

Our approach bridges the gap between network and content analysis in automated communication analysis systems, supporting domain experts in exploring and analyzing arbitrary bi-directional communication. At the same time, we aim to pave the path for a more holistic approach to communication analysis.

## 2. Related Work

The **origins** of communication analysis can be traced back to the works of Simmel [Sim08], studying interaction in sociology, and Moreno [Mor34], researching human networks and laying the foundation for social network analysis. This technique, which describes a collection of research methods for identifying structures in systems, is widely described in the standard literature [Sco17] and applied in many fields. For **communication analysis**, the early works were later extended and taxonomized by Bavelas [Bav50] and Leavitt [Lea51], before Roger [RK80] proposed to extend the field to communication networks. While detailed, domains-specific content analysis, for example, in psychology [Ber52], was already known almost seven decades ago; only the recent advancement of computational capabilities allowed the focus to shift to a bulk analysis of communication data on a larger scale. Using methods from social network analysis, it became possible to investigate **network aspects**

like social ties and communication behavior [LZ15] by using centrality measures, detect communities [XKS13] and clusters [AW10] or model whole artificial networks [BMBL09]. However, using social network analysis on communication data primarily covers these network aspects. When focusing on **metadata and communication content**, a virtually unlimited amount of analysis methods can be applied. For example, metadata analysis [dMHVB13] can be used to identify individuals, keyword-based searches [YP04] can filter for specific content, while methods from natural language processing [MS99] like sentiment analysis [GH11, PL08], topic modeling [RS10], or lexical chaining [GREA15] can support an advanced understanding of the meaning.

While this should give us ample scientific and technical methods at hand to analyze communication thoroughly, when we study automated, **digital human communication analysis systems**, we notice the peculiar oddity that most existing approaches focus either on the network or the content aspect, but not both, as we discussed above. A majority of the systems with communication analysis in mind follows the former approach. The de-facto standard in civil security and business intelligence applications are IBM's i2 Analyst's Notebook [IBM20], and, to a lesser degree, the large network analyzer Pajek [BM98], both **commercial solutions**. The open-source equivalent Gephi [BHJ09] is also used sometimes. While i2 Analyst's Notebook can be extended with content analysis capabilities, such a search is only offered as a separate interface. From a **visualization perspective**, all follow a node-link-diagram based approach. These suffer from inherent limitations like clutter or occlusion when the graph size becomes too large, and connections cannot be filtered enough for the search tasks, while techniques like edge bundling can only help so much. In fact, a study by Ghoniem et al. [GFC05] shows that matrix-based visualizations are better suited for large or dense networks and perform better from a scalability viewpoint. Various other methods are described in a general survey [SSG12] of visualization systems for large networks. For example, when considering communication networks as multivariate graphs, one could employ techniques like Multilevel Matrices [vH03], LiveRAC [MMKN08], Hyper-Matrix [FAS\*20] or Responsive Matrix Cells [HBS\*20] for improved scalability and detail, often combined with matrix reordering techniques [BBH\*16]. Alternatively, it is possible to leverage semantic or magic lenses [GSBO14] to highlight and enlarge relevant parts. When focusing on social network exploration, the survey by Henry Riche et al. [HF10] focuses on specific extension for node-link and matrix-based approaches.

Looking at the **academic contributions**, we discover mostly alternative visualization and analysis methods. **Matrix-like** approaches are used, for example, by GestaltMatrix [BN11] to visually analyze asymmetric relations, or MatrixWave [ZLD\*15] for comparing multiple event sequences. A notable set of approach that leverages matrix designs were proposed by Nathalie Henry: MatrixExplorer [HF06] presents the idea of combining node-link and matrix approaches, which NodeTrix [HFM07] extends to address the occlusion problem for large node-link diagrams by switching to matrix view locally. To address issues in path tracing in matrix views, they further present MatLink [HF07]. **Timeline-based** designs were proposed as part of Timeline Edges [Rei10] to efficiently use edge space, in T-Cal [FZC\*18] to highlight areas with high communication volumes using distorted plot lines or as part of CloudLines [KBK11] to

display event episodes in multiple time-series. **Hybrid approaches** also exists, like Fu et al. [FHN\*07] that propose to modify graph representations using multiple planes to recognizing communication patterns in e-mail networks. When considering the metadata and content analysis side, countless methods exist in various fields. However, many do not explicitly focus on communication analysis, and we will not discuss them here, although some can in principle be applied to a selection of the content-related tasks defined above (e.g., the interactive discourse analysis by Zhao et al. [ZCCB12]).

Leveraging analytical capabilities from both network and content information **simultaneously** has rarely been done. Interestingly, **commercial systems** seem to be ahead of their academic counterparts. Apart from Analyst's Notebook, which we discussed above, systems like Nuix Discover and Nuix Investigate [Nui20] for e-mail analysis and whole investigation frameworks like Palantir Gotham [Pal20] and more recently, DataWalk [Dat20] have become available. Some have received mixed responses by the public given their primary application in the intelligence and law enforcement community. As they are commercially developed, closed source solutions with few details about their detailed capabilities and internal workings, as well as their applications, they are often shrouded in secrecy (given their target domain). This proves problematic because it hinders oversight from an independent community like academia to track capabilities or point out issues like bias, which becomes increasingly relevant with the usage of machine learning techniques within these solutions. Looking at the **academic contributions**, we have three relevant approaches which combine network and content perspectives: TimeMatrix [YEL10] by Yi et al. combines meta-information and network structure. It uses a matrix-based visualization to analyze temporal social networks using TimeCells, showing a visual aggregate of a node's temporal information. For example, it can show edge count for a pair of nodes over a period of time. OpinionFlow [WLY\*14] by Wu et al. combines content with network structure analysis to visually analyze opinion diffusion. They base their modeling on the network structure and sentiment-specific word embeddings, with the results using timeline-like visualizations. IEFAP [HDL\*09] by Hadjidj et al. also combines content with network structure analysis. It uses a multiple-coordinated view system with a node-link diagram as the primary visualization to support the forensic e-mail analysis, supporting various filter techniques, like metadata or keyword analysis and authorship attribution.

Given the little overlap between these solutions, their restricted applicability to communication in a generic case, and the growing support in commercial applications - in contrast to academic literature - reveals a missed opportunity. This is the **gap** we aim to fill: Provide a blueprint for a more holistic approach to communication analysis that supports network, metadata as well as content aspects simultaneously by the use of extendable plug-in models in a single interactive visualization system, enabling the effective exploration of communication for interrelated tasks as defined above.

### 3. Challenges and Design Decisions

Such an approach encounters several challenges. One obstacle comes in the form of the different requirements to internal data representation and the **analysis methods**, like graph-based approaches or content-based methods as well as the communication type involved,

and how they can be combined in a single system while acting on the same data set. For this, we formalize communication modeling in Section 4 and describe the analysis as abstract operators working on a shared data space used internally.

The second challenge concerns the **visual representation and interaction** when combining these different methods in a single framework. The proposed system has to visually support different analysis modalities as part of a holistic framework through understandable and effective visualization methods, provide easy access to the visual results, and to allow useful interactions between them. The visual design choice also depends on the size and sparseness of the communication network. For example, a design based primarily on **node-link diagrams** like IEFAF might work well for very small networks, but requires larger ones to be sparse or decomposable into interrelated communities. However, choosing node-link diagrams makes it very hard to integrate additional information in an accessible way [Rei10], with coloring, overlays, and details on demand as options. An alternative [GFC05] is to use **matrix-based approaches** to support larger and denser networks, which also support in-cell content. We will follow this path for our approach. Compared to TimeMatrix, we extend the matrix-based approach considerably by using multiple views involving semantic zooming within the matrix visualization and thereby displaying specialized visualizations in-line and on-demand. Our approach scales well with the number of messages, which, as edges, are the primary source of clutter in a node-link-diagram. However, when the number of users exceeds several hundred [FAS\*20], options like scrolling or magic lenses might be required. Our design focuses on a holistic approach, in contrast to existing approaches with limited exploration concepts or heavy task adaption. It offers flexibility when analysis tasks require combining methods from different sub-fields of communication analysis. However, when analysis tasks are very specialized, for example exploring the network structure alone, a node-link approach might be more suitable. Further, a possibly viable addition to our design choice would be to use **coordinated views**, providing spatially separate visualizations that are logically linked. Such an approach could be explored further (see Section 7).

#### 4. Modeling Communication

In the following two sections, we describe the overall workflow, shown later in Figure 2. We begin by defining requirements for an abstract analysis level in Section 4.1 and define standard task levels to address common functionality. For the purpose of this work, a level can be thought of as a module which answers a individual analysis aspect. A level can have both a view (see Section 5.1) to show interactive visualizations and properties (see Section 5.2) to configure it. We give two exemplary descriptions of more complex, extendable levels as a blueprint for individual communication analysis. Firstly, a dynamics level in Section 4.3 to analyze network and metadata. Secondly, a level for semantic concepts described in Section 4.4. The architecture of the framework makes it easy to add own levels. While it is desirable for levels to cover distinct analysis aspects, they are not restricted from covering overlapping aspects. In Section 5, we then discuss the integration of the individual levels and the interactive exploration using visual analytic principles.

#### 4.1. Abstract Level

A communication network can be described as a multidigraph  $G := (V, M)$ , with  $V$  a set of vertices representing the communication participants and  $M$  a multiset of ordered pairs of vertices representing a communication event. Additional metadata and content can be modeled by defining an information function  $i : M \rightarrow D$ , mapping a communication event to a data space  $D$ . Individual analysis levels can now be generically defined as operators that act on the vertex space  $V$ , edge space  $M$ , and the information function  $i$ . All this together forms the graph-like shared data space, which is used internally to store all information. Hashmaps and support index structures are used for efficient access. Each level can have none, one, or multiple in-line visualizations called views in the main interface (see Section 5.1). These visualizations can transfer domain- and task-specific information relevant to a domain expert. Further, each level has its parameters and filters for control. As individual, separate levels itself would not provide many benefits, the key idea is to complement each other on the system scale. Their flexible and simultaneous application in a single approach provides support for cross-matches, as level-specific filtering adds together to form a global filter. Additionally, all levels can output a feature vector that is fed to a machine-learning-based retrieval system, described in Section 5.3, to enable intelligent user steering. The system can be customized and extended to more specialized tasks by adding additional levels to cover more specific needs.

#### 4.2. Standard Task Levels

Analysis usually requires a set of standard operators for filtering and selection, so we provide a set of standard task levels. For example, to support simple tasks like restricting the time ranges, one can define an operator on  $M$  which filters edges based on the timestamp information in the data space  $D$  (a **timefilter level**). Other examples are to filter participants in  $V$  through properties in  $M$  and  $D$  (a user **selection level**), or keyword-based search by restricting based on content information in  $D$  (a **keyword search level**). As these levels act primarily as filters, a corresponding view (see Section 5.1) might not be required. To provide basic visual analysis, one can define an operator on  $V$  and  $M$  (a **volume level**) which tracks the amount of communication between users, or an operator on  $V$ ,  $M$ , and  $D$ , that track the temporal evolution of such communication (a **distribution level**). For both, we provide corresponding views in the main interface (see Section 5.1). For the remainder of this work, we will primarily focus on the more complex levels in the following two sections as they allow us to define task-specific analyses and only consider these standard task levels when necessary.

#### 4.3. Dynamics Level

Different questions are of interest when analyzing the communication behavior between entities: For example, how does the volume of communication develop? Is communication discontinued? Is it one-sided, or are there specific patterns? However, if we look at communications only as individual messages, it may be difficult to answer such questions. To analyze the dynamics of communication events more thoroughly, we follow our previous work [SFS\*19], and define a set of features which operate on the edge space  $M$  and

the information function  $i$ . There, we model communication not only as individual events but as a flow, which can be described using distributions and a continuous density function. This view enables us to easily model influence effects like the response time (both prior and delay), the width of the temporal influence, or control between spikes and general tendencies simply by adapting parameters like  $\mu$ ,  $\sigma$ , or  $h$ , respectively. For details, see the original publication. It further allows us to detect breaks in communication and thereby identify *communication episodes*. The choice for the granularity of the episodes is made globally depending on the dataset. Together, this facilitates the structural communication analysis and helps to address the questions posed above.

In our work, we leverage these previous ideas and adapt them to work on top of the abstract level formalization defined above: We take the properties and visualization ideas from the original work and transform as well adapt them to work with this approach by adding level properties and an in-line visualization (a view), which we further describe in Section 5.1.






#### 4.4. Thematic Level

In general, the inclusion of thematic concepts allows a user to refine their search task in a more powerful way than keyword lists and comes more naturally to analysts, who often think in concepts. Regarding modeling, a thematic level operates on the edge space  $M$ , depending on the content information in the data space  $D$ . A standard method to extract concepts from text data is named entity recognition. There, it is possible to either use pre-trained models or adapt them with domain-specific or task-specific concepts. However, a simple search using these semantic concepts might not be flexible enough to allow for more complex search tasks like "retrieve communication talking about a person in connection with a location". Therefore, we propose an interactive visual query language that allows for a flexible combination of semantic concepts to fulfill such search tasks. This query language allows creating multiple semantic queries based on spatial co-occurrence of semantic concepts. These queries can then be combined using Boolean algebra to build more complex filters. For example, the above search could be restricted further by additionally requiring an organization to be mentioned, which is combined with the first query. As this level acts more like a filter on the data, it is an example of a view-less level, not having a separate in-line visualization in Section 5.1.


### 5. Visual Interactive Investigation of Communication

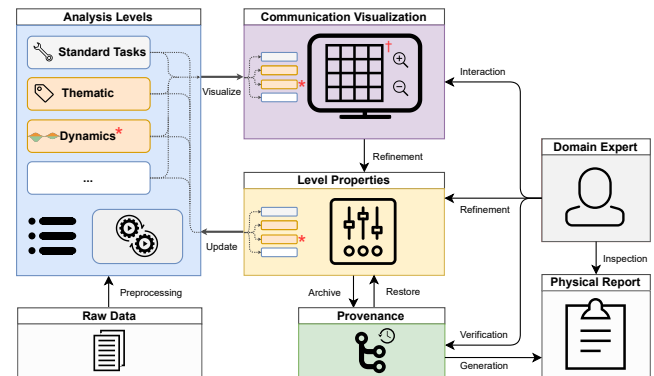
This section focuses on the visualization and interaction concepts to integrate multiple levels in a single framework while providing a tight coupling between the network and content analysis aspects. The proposed workflow for this architecture is described in Figure 2. We begin by describing how the overall network can be visualized using a matrix-based visualization that provides multiple levels as views, representing the individual analysis levels' results. Table 1 shows the interplay between Levels, Views, and their properties.









Conceptually, the information becomes more nuanced during level drill-down, going from overview to specific analysis to content, while each level addresses a specific question related to the level-modeling in Section 4. To facilitate the interactive exploration,

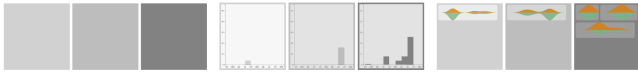
Level	View	Properties
Volume	X	-
Distribution	X	-
Timefilter 	-	X
User Selection 	-	X
Keyword Search 	-	X
<b>Thematic</b> 	-	X
<b>Dynamics</b> 	X	X

**Table 1:** The different analysis levels in our system. Among the standard tasks levels, the two examples of more complex, custom level implementations are highlighted in bold. Components indicated with \* [SFS\* 19] are based on previous work.

the levels can be controlled via a property pane . The levels can then act as filter methods, enabled through standard operators for standard task levels, steering options for conversational dynamics, and a visual query interface for thematic searches. We specify how these individual levels can act together, which in-line visualizations they provide to support the exploratory analysis, and how user feedback for the adaptable retrieval system can support the search. This methodology helps domain experts to gain a better understanding of the communication data by providing rapid-feedback through interactive filtering, covering different analysis levels simultaneously. Finally, we describe how all steps are recorded in a provenance history graph, making the decision-making processes traceable.



**Figure 2:** High-level workflow of our system, highlighting the main components and the interaction flow for the communication network analysis. The workflow begins with raw data  extraction and the generation of the individual level . A multi-level matrix visualization  containing other in-line visualizations presents the current model state to the user and allows for different interaction and exploration schemes. The domain expert  can then explore and refine the levels through their properties , leveraging an internal relevance feedback system , updating the overall model state, and adapting the selection. The history of refinements is archived to provide provenance information , which can be exported as part of a physical report  for inspection, traceability, and explainability. Components indicated with a \* [SFS\* 19] and † [FAS\* 20] are based on previous work.



(a) *Volume View*: The amount of communication is visually mapped to the color of the cell. (b) *Distribution View*: Provides an overview of the temporal distribution using barcharts. (c) *Dynamics View*: Visualizes the communication episodes between two entities.

**Figure 3:** Overview of the three views we offer. We distinguish between the volume, distribution, and dynamics view. The first two in-line visualizations come from standard task levels (see Section 4.2), the latter from the dynamics level (Section 4.3).

### 5.1. CommAID Interface Design

For the visualization of the communication networks, we adapt the multi-level matrix technique from our previous work [FAS\*20]. However, we change the meaning of grid layout as well as the levels and the cell information: (1) Instead of representing nodes (rows) and hyperedges (column), they, respectively, become senders and receiver. (2) Instead of displaying increasingly detailed cell information, the in-line visualization represents the results of the individual levels discussed in Section 4 as independent views, which are different, but not necessarily more detailed. The overall interface of our approach is shown in Figure 1. The novelty here is how existing visualization are combined, adapted, and integrated into an holistic framework and how the interaction with it is designed. Apart from the interactive matrix-based visualization, the linked level property pane allows to restricting the search space, using standard task filters, dynamics settings, and a thematic concept builder. In this prototype, three different views are provided: *Volume*, *Distribution*, and *Dynamics*. They are shown as part of Figure 1 and in more detail for three generic cells each in Figure 3. A provenance history graph, discussed later in Section 5.4, allows to keep track of the analysis steps and results. In the following, we explain the design rationale of the views. It is important to note that through semantic zoom, the order in which the views are shown is fixed.

The basic principle of semantic zoom is that each cell of the matrix visualization serves as a canvas for a different type of analysis result of the communication between two entities in the network. However, rendering detailed visualizations there makes sense if a cell has a specific minimum window size. Otherwise, even basic visualizations can be impossible to read. Guidelines [FIBK17] have been developed for the required size to retain readability. Along those lines, and with the type of views in mind, we have chosen a view switch with every doubling of cell size. When using a different type of view, the transition criteria might have to be adapted. For example, either by using a different scaling factor or keeping the cell size for some view transitions and just switching the view.

The **Volume View** (belonging to the volume level) displays the number of communications between two entities, where the amount of communication is visually mapped to the cell’s color. Different color scales can be used depending on the task requirements. Figure 3a shows a sequential, single-hue gray color scale, where white indicates that no communications are taking place, and black represents the maximum number of communications between two

entities in the network. Color schemes are replaceable, for example, for users with visual impairments or by using diverging color schemes to indicate deviations from the average.

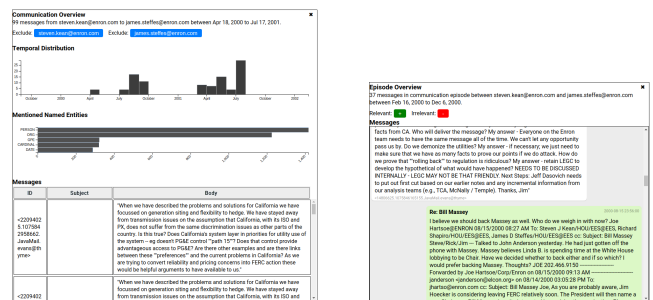
The **Distribution View** (belonging to the distribution level) is used to provide an overview of the temporal distribution of communications. Similar to TimeMatrix, we use a bar chart, but add a background color (matching the Volume View) to retaining the overall amount information. Thus, in addition to the temporal distribution of communications, global information can also be visualized.

While these two represent views for the standard task levels, providing views for the custom models is especially interesting. Here we offer the **Dynamics View**, visualizing results from the dynamics level. There, we represent the communication episodes between two entities in the network. Depending on the tasks, the episodes can be shown chronologically or customly sorted.

All three views have in common that they offer additional details-on-demand. A click on a cell opens a zoom-level-dependent tooltip (see Figure 4a), which provides information about the time distribution, named entities used, or raw data. A click on an episode also opens a tooltip (see Figure 4b), which visualizes the discussion content between two entities using a chat-style metaphor. In both details-on-demand visualizations, the user can directly perform a refinement step, e.g., by excluding entities from the search or evaluating communication episodes for relevance.

### 5.2. Level Properties

Each level can have its own properties, accessible through a property pane on the right in the main user interface. The standard task levels offer controls like cutoff values, include/exclude lists, or time sliders. The dynamics level offers restrictions on the individual properties of conversational dynamics [SFS\*19]. Here, we want to describe one more complex property that can be offered on the property pane for custom levels, using the **Thematic Level** as an example: a visual query interface for thematic searches using named



(a) *Details-on-demand populated by the distribution and thematic level presenting the time distribution, the thematic named entities used, and, additionally, the raw data.*

(b) *Details-on-demand for a communication episode visualizes the discussion in a chat-style metaphor. The communication can be ranked for the classifier discussed in Section 5.3.*

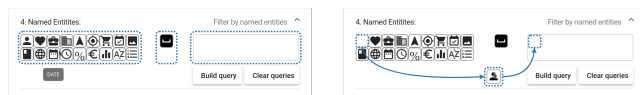
**Figure 4:** Overview of the details-on-demand offered by different semantic zoom levels, provided by different views.

entities. To generate the named entities, we employ a pre-trained model from spaCy [Hon19], containing a set of 18 named entity categories. The interface itself is shown in Figure 5, illustrating the individual components and the step-by-step process for creating a sample named entity relation pattern query, searching for two concepts that occur within a specified word distance.

As shown in Figure 5a, our visual query interface consists of three main components: A repository containing a set named entities, such as persons, appointments, or organizations, a special token to allow distances between naming entities, and finally, a query building component. To build a named entity pattern query, a user can drag one or multiple individual named entities from the repository and (optionally) the special token into the query building component, as highlighted in Figure 5b. The concepts can also be rearranged inside the query building component, reflecting different query types, like single concepts, a chronology order of concepts, or distances between the concepts. For the latter, the maximum allowed word distance between two named entities can be set directly in the corresponding token in the query builder. In the example shown in Figure 5c, the maximum allowed distance between named entities is set to seven words. After all necessary corrections and adjustments have been made, the current query can be added to the set of applied queries (Figure 5d). The query shown only serves as a single example of a named entity query, where other types are possible.

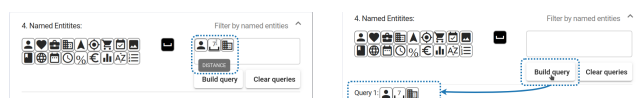
### 5.3. Machine-Learning-based Retrieval System

The design of using multiple complementing levels allows for cross-level search, with level-specific properties that can act as filters, acting together to form a global filter. This already allows for more powerful and interrelated search tasks than the application of individual levels alone. However, depending on the types of filtering defined and their interactions, the possible settings might overwhelm domain experts. Therefore, we propose that each level can output a feature vector for a communication event. Level-specific vectors



(a) The named-entity repository as icons, a distance token and the query building component (f.l.t.r).

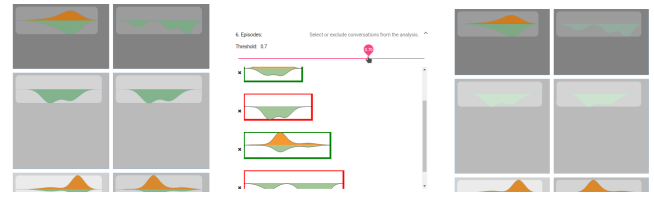
(b) Building a named entity pattern query is performed via drag&drop of icons and token.



(c) The concepts can be interactively rearranged, reflecting different query types. The maximum allowed word distance between two named entities can be set directly in the corresponding token

(d) After all necessary adjustments have been made, the current query can be added to the set of applied queries. Using conjunctive (AND) or disjunctive (OR) combinations multiple queries can be combined.

**Figure 5:** Illustration for the step-by-step process for creating a named entity relation pattern query, searching for two concepts that occur within a given word length of each other.



(a) Cutout showing the individual episodes. (b) The positive and negative samples. (c) Results, with irrelevant episodes faded.

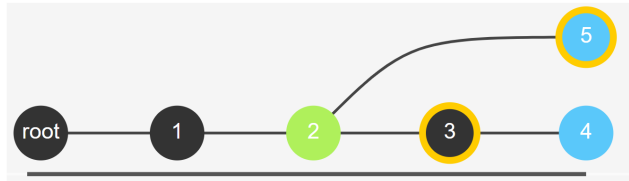
**Figure 6:** By providing feedback, users train ML models to identify relevant conversational dynamics in episodes. Here, the aim is to identify episodes in which the selected groups start the conversation, leading to a discussion of both entities.

can be combined to a single, large feature vector used for classification purposes in a user-steerable machine-learning level. Although progress has been made to use deep learning efficiently by reducing training time [SK16] and improve explainability [GMT\*18], their usage is still problematic when requiring (theoretical) traceability, for example, due to legal constraints. Consequently, as proposed in the literature [MQB19], we employ a rule-based approach based on a random forest model. However, while this can fulfill legal requirements, from a perspective of lay use, a random forest's decisions might still be tough to understand. It has, however, the additional benefit that the training size can be relatively small (usually much less than a few dozens), make it suitable for an interactive application, while the examples can easily be collected by the users themselves, based on their expert knowledge. A user can train individual automatic classifiers that support him on specific tasks and modularly combine them to overall predictions.

The selection of training examples for the classifier happens interactively. A user can label communication binary as relevant or irrelevant to perform a semi-automatic classification of communication into user-defined classes. An example of such a selection is shown in Figure 6. Such a trained classifier can then perform the binary classification for all other communications, acting as an additional high-level filter. Since we use a Random Forest Classifier, we can model the uncertainty for the prediction, which is useful for threshold. This also allows presenting ambiguous communication, where the classifier is very uncertain, to the user for re-labeling, allowing for an interactive optimization. To separate between this semi-automatic retrieval system and manual level property settings, communication that is filtered out by the automatic system is only faded out based on a variable threshold, but not hidden completely.

### 5.4. Analytical Provenance

To support experts even in lengthy and complicated investigations, we offer a provenance history component, which is displayed in Figure 7. Since explainability is relevant, for example, in court cases, decision making and a record of the obtained results must be preserved. The provenance history contains previous filter states as nodes in a directed acyclic graph (DAG) to allow for more flexible navigation compared to a linear history [HSS\*18]. Important states can be visually starred. The user can navigate between different states, go back to previous results, or branch off as a new starting

**Provenance History:**

**Figure 7:** Provenance history component showing previous filter states as nodes in a directed acyclic graph (DAG). The currently selected node and the leaf nodes are specially colored and important states can be starred for easier distinction. The states allow for more flexible navigation, to revert from dead ends or branch of to a different analysis direction.

point for further analyses. This enables the user to continuously verify and retrace results, which is especially advantageous for creating trust in the user's results. Further, the individual filter states can be bundled into a physical report and thus archived. In this way, the analyst's results can be reproduced, retraced, and explained, even after an extended period and/or to third parties. Since each of the steps and all of its belonging information and metadata and the obtained results can be reviewed and analyzed independently, this allows for explainability of the results obtained.

## 6. Domain Expert Assessment

To demonstrate the effectiveness and improvements compared to existing approaches of the visual exploration of communication behavior in COMMAID, we conduct an expert assessment of the prototype while additionally conducting a case study together with a small-scale user study later on. As communication data, we use the largest publicly available source, the Enron dataset [KY04], encompassing 517 431 messages from 151 users.

### 6.1. Formative Expert Assessment

The assessments were conducted by demonstrating the prototype to six domain experts (LEA 1–3, RS 1, SI 1–2).

**Expertise** LEA 1 is a criminal investigator at a European law enforcement agency with extensive experience in the field, including communication and network analysis with graph-based visualization using systems like IBM i2 Analyst's Notebook, Pajek, and Gephi. LEA 2 is a criminal investigator with no prior experience using graph-based visualizations. He focuses on the communication content, using forensic tools like Cellebrite and IBM i2 Analyst's Notebook, which is laborious. LEA 3 is a senior judicial commissioner in law enforcement with extensive experience in digital investigation techniques. He is aware of the graph-based approaches used within his unit, but limited own experience. RS 1 is the head of a university-affiliated institute for policy and security research, a full professor and senior researcher working on government projects. He has worked with graph-based visualizations and approaches for over 20 years, for instance, for bibliometric investigations. SI 1 is a senior project lead within the security industry for developing investigative solutions for LEAs. SI 2 is a junior solutions specialist

within the security industry and worked on visualization techniques, including graph-based visualizations, for criminal investigations before. All but one (SI 2) of the experts each have more than 15 years of experience in digital and criminal investigations.

**Methodology** The expert assessment was conducted as a formative evaluation taking 90 minutes, with the experts observing, commenting, and asking questions on the system, while provided with a user sheet. They were also allowed to steer the exploration by requesting specific actions. As such, a formative evaluation without direct usage usually cannot replace the benefits of a full user study; therefore, we additionally conducted a user study with two further domain experts later on, as described in Section 6.2. The experts were first given a ten-minute introduction about the aim, which is aligned with the tasks defined in Section 1: facilitating the visual analysis and exploration of large amounts of communication data using different visualization and filtering methodologies simultaneously to structure and reduce the search space to a manageable size for inspection by an analyst. The prototype was then showcased and explained during 30 minutes of interactive demonstrations, where the experts actively asked questions and commented on the presented aspects. They include using the network overview to detect promising connections and explore individual details as well as communication episodes in level-specific visualization through drill-down. Further, we presented and then debated the different interaction techniques, then the available filters to reduce the selection, and finally the example-based machine learning retrieval classifier.

The interactive session was followed by a structured interview (appendix), taking about 50 minutes, using a set of 29 prepared questions about various aspects of the approach. This interview was interluded by interactively presenting aspects in the demonstrator when requested by the experts. The session aimed to elicit experts' opinions on the system's design and interaction decisions and identify aspects they find helpful or prone to misinterpretation. Further, we were interested in how they would apply these methods in their specific workflows and criminal investigations in general. The findings of these assessments are described in the following.

**Findings** All the experts state that both the approach of using a **matrix-based overview visualization** and using a semantic zoom for more details is a new approach in their domain. For example, according to LEA 1, he has "always worked with graph tools" so far and thinks of our technique as "very interesting and helpful". All experts think that a matrix-based visualization is superior to graph-based approaches in "terms of scalability" (SI 2) and displaying "supporting information" (RS 1). However, both LEA 1 and LEA 2 recommended the matrix columns to be freely reorderable.

Regarding the **semantic zooming**, the experts are familiar with such a concept from everyday applications like digital maps. For communication analysis, some did not expect this functionality at first (cf. LEA 2). However, it supports their work and is an excellent way to drill down to "go into the raw data" (cf. LEA 2). RS 1 judged the semantic zoom as intuitive, but expected that - instead of the communication structure and content - more information about the "importance of relations" (RS 1) is shown. RS 1 proposes to include such information as another level, as the design is "flexible enough" (RS 1). For example, an **analysis levels** for centrality analysis could be added.



In terms of **filtering**, the experts are happy to have the standard task functionality included. However, more advanced concepts like the semantic named entity search were “unexpected” (LEA 1, LEA 3). In their previous experience, the LEA’s were only able to search using lists of keywords and were “never able to search for concepts” (LEA 1). They regard this functionality to have much potential, as it allows for “more generic” (LEA 1) and high-level search terms. As the prototype’s current implementation restricts them to AND-based queries, all experts state they would like to combine the queries more freely using Boolean logic.

Regarding the **machine-learning based retrieval** of communication episodes, all experts agreed that detecting related and sequential communication is “important for contextual information” (SI 1). The visualization as density-based communication amount can be intuitively understood by all experts. LEA 2 regarded it as beneficial that the detailed raw data from a communication episode can be inspected within the visualization. Going on from there, the ability to train a machine learning model by giving communication episodes as examples is viewed as “opening up new possibilities” (RS 1). Domain experts especially favored that **arbitrary features** can, in principle, be used for the machine learning model (extending upon those we defined above), which can be communication “based on text, audio, pictures, geographic information systems, or combined with graphs” (RS 1), using additional levels. Therefore, the features itself “do not matter much, as the user has to define them” (RS 1), making the flexibility of our approach “very broad” (RS 1).

With such broad applicability, the explainability and retracing of the results are “an important issue. If an analyst has a result, he needs to explain how he ended up there” (LEA 3). This explanation is simplified tremendously “with the generation of a step-by-step report” (LEA 3), as producible from our **provenance** history, whereas currently, “analysts have to write detailed accounts on how they got to the information” (LEA 3) and justify it each time in writing.

In terms of practical usage, they had “no ideas for additional [conceptual] features” (LEA 1), except for the inclusion of centrality measures. But the potential possibilities with the framework are almost “overwhelming at first” (SI 1). The **applicability** of the presented approach is not restricted to a narrow use case as the one presented. Therefore, the system is “broadly applicable to multiple domains where you have bigger groups of communication data. Data that [law enforcement experts] often have to deal with. For example, organized crime, financial crime, or terrorism” (LEA 3). All the other experts share this view. Indeed, the presented system is “a beginning of an interaction platform where you can combine other logic’s as well and [which] offers many possibilities” (RS 1), providing custom analysis levels specific for your needs (cf. RS 1).

## 6.2. Case Study and User Evaluation

During the formative expert assessment, the experts did not interact on their own with the system, as we were only able to secure a limited, non-individual amount of time with them. To compensate for bias, we additionally conducted a small-scale user study with two further domain experts (LEA 4 and RS 2). For this, we describe a possible case study, highlighting the benefits of an holistic approach.

**Expertise** LEA 4 is the head of the big data department of a

federal governmental agency supporting criminal investigations. She advises law enforcement agents on the applications of artificial intelligence to criminal investigations. RS 2 is a senior scientist at a federal government research institute with over ten years of experience in communication analysis and terrorist investigations.

**Methodology and Case Study** The case study is based on an artificially financial fraud use-case to identify **senders**→ and **receivers**← of relevant communications. The aim is to discover those persons 👤 which, during the **first nine months of 2001** 📅, disseminated knowledge about **legal** 📄 issues involving **persons** 👤 in combination with **organizations** 🏢 in **California**, and then identify the only person 👤 who received information from all of them. The experts were given a system manual and a short written description of this task. Such a task often occurs in real cases, but is not well supported by existing approaches. Instead, one often relies on keyword list using domain-specific ontologies, requiring manual work to create and search through the results. Here, we present how an **holistic analysis** (see also appendix demo video) of **network structure and dynamics** 📊, **metadata** 📄, **keyword-based search** 🔍, and **semantic concepts** 🗨️ can address this task.

The experts were expected to map the task conditions to the analysis levels and their properties: namely **specific time ranges**, applicable **field offices**, and concepts that are relevant for the search ( **persons** 👤, **organizations** 🏢, and **legal topics** 📄 ). In a second step, when they have identified the persons 👤 that disseminated the knowledge, they were expected to make the mental transfer to restrict the view to those participants as **sending persons**→ and examine which participant is the only **receiver**←. The successful completion of the task was measured by checking the identified users. During the task, the users were undisturbed to explore and try out the system freely. They themselves could decide how long they want to train and check out the system, before actually starting. It was the first time they saw and used the system for both domain experts, having no prior experience with it and having received no introduction, so we could do a blind test. After completion, we interviewed them shortly about their experience.

**Findings** Both experts were able to successfully complete the task within 15 minutes. The interface and **interaction concepts** were described as “intuitive [and] self-explanatory” (LEA 4). The first expert was surprised to be offered a search for concepts and initially tried to use the keyword-based search level for conceptual searches instead of the visual query interface, while the other did not make this mistake. We conclude that some users would benefit from a more hands-on explanation. LEA 4 took five minutes longer to complete, exploring the options and results in between, but ultimately solved it like RS 2 (see solution above), who went ahead directly. The system offers “helpful” (RS 2) drill-downs of visualizations and is intuitive and straightforward (cf. RS 2) to use. Both domain experts noted that the system naturally supports the investigative **workflows**, and the interaction design combined with the documentation is sufficient for working productively and getting relevant results. Compared to their existing systems and workflows, the system provides a significant benefit in analytical capabilities. Most notably, it allows the **simultaneous application** of different search methodologies to support cross-matches. This allows for more powerful queries, in contrast to manually merging separate results.

## 7. Discussion and Future Work

During the formative expert assessment and the study, we received several proposals on how to extend our approach further. Leaving out expected requests for a research prototype like more supported data imports, we instead focus on the core functionality of the approach by discussing the limitations and the broader applicability as well as the context of future work. For our prototype approach, we adapted the generic blueprint of modularized communication analysis to the case study by providing two example levels.

Of course, the system can be extended modularly with **further analysis levels**, for example, those featuring graph centrality measures, community detection, leveraging specific meta-data like locations, or more specific content analysis modules based on linguistics. Currently, this requires modification in the source code, but a plug-in architecture would be conceivable. Regarding the overall **matrix visualization**, the challenges and limits of such an approach for communication analysis have partly been discussed in previous work [FAS\*20]. When tasks are primarily focused on network structure analysis, a classical node-link-diagram-based approach might be more suited. Alternatively, one could consider a coordinated view approach and adding a synced node-link-diagram. As space is limited, making both matrix and node-link components user resize-able would allow for a task-adaptable interface. Intelligent layouting of the node-link diagram based on different parameters (centrality, connectedness, meta-data) would support visually finding patterns, which can be further analyzed in the matrix view and vice versa.

A challenging step for future work is to investigate how this approach can be used for the analysis of **multi-party conversations**. So far we duplicate messages with multiple recipients, which partly destroys the group aspect. Future work could investigate how hypergraphs allow to capture such scenarios or how additional zoom level could be leveraged for group communication analysis. Also, we assumed some practical restrictions to describe the network by requiring receiver, sender and timestamp information for each message. As some analysis steps requires those, incomplete entries (like unknown recipients) cannot be represented. Another aspect of the prototype is the **visual query language**, where it could be valuable to extend the grammar and support nested queries visually.

Finally, the expert interviews and the user study resulted in a positive response to our workflow and prototype system. However, we are well aware that the sample group for our formative experts' assessment and the user **study was quite limited**. To achieve statistically more accurate results and broaden the perspective, the study could be extended with more participants as part of future work. This explainable and accountable decision making is not only relevant in the security domain, in which the case study and expert assessment were conducted. Indeed, the experts think the applications are not limited to such a narrow set of criminal communication investigations but can be applied to communication data in **other domains**. One different application would be in the business intelligence domain. The system could be applied as a search and retrieval mechanism to search for hidden, decentralized knowledge contained in business documents and communications. This knowledge can then be linked and extracted into centralized knowledge management systems, allowing for more efficient management structures and avoiding redundancies, making the processes more accountable.

## 8. Conclusion

So far, most interactive, automated communication analysis approaches focus either on the network aspects or on the content, in contradiction to communication research. As such, the individual or isolated analysis may not suffice to capture the full available information and may lead to less effective, incomplete, and biased results. Further, it can increase the struggle experts face when articulating their domain knowledge, not leveraging their full potential.

We address this challenge by arguing for and discussing a holistic approach to communication analysis, simultaneously applying both methods, allowing for more structured and detailed analytical capabilities. To help domain experts deal with the complexity of modern communication data, we present **COMMAID**, a blueprint for a visual analytics-based communication analysis system that offers a wider approach, providing a tight coupling between the network and the content analysis aspects, building on individual levels and supported by a machine learning-based retrieval system.

We provide two extendable levels as an example for network and content analysis each, covering dynamics modeling (based and extend from our previous work [SFS\*19]) and semantic text analysis. We leverage ideas from hypergraph analysis [FAS\*20] for a multi-level matrix-based visualization design to integrate those levels in a single interface. However, we specifically tailor and adapt this idea to communication analysis by providing specific visualization levels to support domain experts in their mental understanding during exploration and allow them to answer more detailed questions about communication behavior and structure, including identifying individual communication episodes. Combining network and content aspects in a single visualization allows for maintaining overview and focus while eliminating demanding context switches, rapidly exploring large search spaces, and providing details on demand. The realized techniques allow the simultaneous analysis of network and content aspects, like properties, conversational dynamics, or conceptual content, to refine the search and supports cross matches.

We evaluate our approach in one case study and through assessments with law enforcement experts using real-world communication data. The results demonstrate that our system surpasses existing solutions, enabling the effective analysis of large amounts of information in a targeted and integrated way. The experts regard this approach as a novel and promising way for a more meaningful communication analysis that can readily be applied to comprehensive analytical tasks as encountered in practical applications. While we focused on communication analysis for law enforcement as driving application, many tasks in communication analysis are similar and, therefore, our methods are more generically applicable to a wider variety of domains, like digital humanities or business intelligence. By bridging this gap between network and content analysis in automated communication analysis systems, we aim to pave the path for a more holistic approach to communication analysis.

## Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) in the framework of PEGASUS under the program "Forschung für die zivile Sicherheit 2018 - 2023" and its announcement "Zivile Sicherheit -

Schutz vor organisierter Kriminalität II". Open access funding enabled and organized by Projekt DEAL. [Correction added on 05 November 2021, after first online publication: Projekt Deal funding statement has been added.]

## References

- [AW10] AGGARWAL C. C., WANG H.: A Survey of Clustering Algorithms for Graph Data. In *Managing and mining graph data*, Aggarwal C. C., Wang H., (Eds.), vol. 40 of *Advances in database systems*, 1386–2944. Springer, New York, 2010, pp. 275–301. doi:10.1007/978-1-4419-6045-0\_9. 3
- [Bav50] BAVELAS A.: Communication Patterns in Task-oriented Groups. *The Journal of the Acoustical Society of America* 22, 6 (1950), 725–730. 2
- [BBH\*16] BEHRISCH M., BACH B., HENRY RICHE N., SCHRECK T., FEKETE J.-D.: Matrix Reordering Methods for Table and Network Visualization. *Computer Graphics Forum* 35, 3 (2016), 693–716. doi:10.1111/cgfm.12935. 3
- [Ber52] BERELSON B.: Content Analysis in Communication Research. 2
- [BHJ09] BASTIAN M., HEYMANN S., JACOMY M.: Gephi: An Open Source Software for Exploring and Manipulating Networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2009), ICWSM, AAAI, pp. 361–362. URL: <http://aaai.org/ocs/index.php/ICWSM/09/paper/view/154>. 3
- [BM98] BATAGELJ V., MRVAR A.: Pajek - Program for Large Network Analysis. *Connections* 21, 2 (1998), 47–57. 3
- [BMBL09] BORGATTI S. P., MEHRA A., BRASS D. J., LABIANCA G.: Network Analysis in the Social Sciences. *Science* 323, 5916 (2009), 892–895. doi:10.1126/science.1165821. 3
- [BN11] BRANDES U., NICK B.: Asymmetric Relations in Longitudinal Social Networks. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2283–2290. doi:10.1109/TVCG.2011.169. 3
- [Dat20] DATAWALK: DataWalk, 2020. URL: <https://datawalk.com/>. 3
- [dMHVB13] DE MONTJOYE Y.-A., HIDALGO C. A., VERLEYSSEN M., BLONDEL V. D.: Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports* 3 (2013), 1376. 3
- [FAS\*20] FISCHER M. T., ARYA D., STREEB D., SEEBACHER D., KEIM D. A., WORRING M.: Visual Analytics for Temporal Hypergraph Model Exploration. *IEEE Transactions on Visualization and Computer Graphics PP* (2020). doi:10.1109/TVCG.2020.3030408. 2, 3, 4, 5, 6, 10
- [FHN\*07] FU X., HONG S.-H., NIKOLOV N. S., SHEN X., WU Y., XU K.: Visualization and Analysis of Email Networks. In *Proceedings of the 6th International Asia-Pacific Symposium on Visualization* (2007), PacificVis, IEEE, pp. 1–8. doi:10.1109/APVIS.2007.329302. 3
- [FIBK17] FUCHS J., ISENBERG P., BEZERIANOS A., KEIM D.: A Systematic Review of Experimental Studies on Data Glyphs. *IEEE Transactions on Visualization and Computer Graphics* 23, 7 (2017), 1863–1879. doi:10.1109/TVCG.2016.2549018. 6
- [FM08] FOLTZ P. W., MARTIN M. J.: Automated Communication Analysis of Teams. *Team Effectiveness in Complex Organizations* (2008). 2
- [FZC\*18] FU S., ZHAO J., CHENG H. F., ZHU H., MARLOW J.: T-Cal: Understanding Team Conversation Data with Calendar-based Visualization. In *Proceedings of the 36th CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), Mandryk R., Hancock M., Perry M., Cox A., (Eds.), CHI, ACM, pp. 1–13. doi:10.1145/3173574.3174074. 3
- [GA08] GURAK L. J., ANTONIJEVIC S.: The Psychology of Blogging. *American Behavioral Scientist* 52, 1 (2008), 60–68. doi:10.1177/0002764208321341. 2
- [GFC05] GHONIEM M., FEKETE J.-D., CASTAGLIOLA P.: On the Readability of Graphs Using Node-link and matrix-based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization* 4, 2 (2005), 114–135. 3, 4
- [GH11] GROH G., HAUFFA J.: Characterizing Social Relations via NLP-based Sentiment Analysis. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011). 3
- [GMT\*18] GUIDOTTI R., MONREALE A., TURINI F., PEDRESCHI D., GIANNOTTI F.: A Survey Of Methods For Explaining Black Box Models. *CoRR abs/1802.01933* (2018). 7
- [GREA15] GOLD V., ROHRDANTZ C., EL-ASSADY M.: Exploratory Text Analysis using Lexical Episode Plots. In *Proceedings of the Eurographics Conference on Visualization (EuroVis)* (2015), EuroVis, The Eurographics Association. doi:10.2312/eurovisshort.20151130. 3
- [GSBO14] GHONIEM M., SHURKHOVETSKYY G., BAHEY A., OT-JACQUES B.: VAFLE: Visual Analytics of Firewall Log Events. In *Visualization and Data Analysis 2014* (2014), Wong P. C., Kao D. L., Hao M. C., Chen C., (Eds.), SPIE Proceedings, SPIE, p. 901704. doi:10.1117/12.2037790. 3
- [HBS\*20] HORAK T., BERGER P., SCHUMANN H., DACHSELT R., TOMINSKI C.: Responsive Matrix Cells: A Focus+Context Approach for Exploring and Editing Multivariate Graphs. *IEEE Transactions on Visualization and Computer Graphics PP* (2020). doi:10.1109/TVCG.2020.3030371. 3
- [HDL\*09] HADJIDJ R., DEBBABI M., LOUNIS H., IQBAL F., SZPORER A., BENREDJEM D.: Towards an Integrated E-mail Forensic Analysis Framework. *Digital Investigation* 5, 3–4 (2009), 124–137. doi:10.1016/j.diin.2009.01.004. 3
- [HF06] HENRY N., FEKETE J.-D.: MatrixExplorer: A Dual-Representation System to Explore Social Networks. *Transactions on Visualization and Computer Graphics* 12, 5 (2006), 677–684. doi:10.1109/TVCG.2006.160. 3
- [HF07] HENRY N., FEKETE J.-D.: MatLink: Enhanced Matrix Visualization for Analyzing Social Networks. In *Human-Computer Interaction – INTERACT 2007*, Baranauskas C., Palanque P., Abascal J., Barbosa S. D. J., (Eds.), vol. 4663 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 288–302. doi:10.1007/978-3-540-74800-7\_24. 3
- [HF10] HENRY RICHE N., FEKETE J.-D.: *Novel Visualizations and Interactions for Social Networks Exploration*, in handbook of social network technologies and applications, borko furht (ed.) ed. Borko Furht. Springer, 2010. 3
- [HFM07] HENRY N., FEKETE J.-D., MCGUFFIN M. J.: NodeTriX: A Hybrid Visualization of Social Networks. *Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1302–1309. doi:10.1109/TVCG.2007.70582. 3
- [Hon19] HONNIBAL M.: SpaCy: An NLP library, 2019. URL: <https://www.spacy.io>. 7
- [HSS\*18] HÄUSSLER J., STEIN M., SEEBACHER D., JANETZKO H., SCHRECK T., KEIM D. A.: Visual Analysis of Urban Traffic Data Based on High-Resolution and High-Dimensional Environmental Sensor Data. In *EnvirVis 2018: Workshop on Visualisation in Environmental Sciences* (2018). 7
- [IBM20] IBM: i2 Analyst’s Notebook, 2020. URL: <https://www.ibm.com/us-en/marketplace/analysts-notebook>. 3
- [KBK11] KRSTAJIĆ M., BERTINI E., KEIM D. A.: CloudLines: compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2432–2439. doi:10.1109/TVCG.2011.179. 3
- [KY04] KLIMT B., YANG Y.: The Enron Corpus: A New Dataset for Email Classification Research. In *Proceedings of the 15th European Conference on Machine Learning* (Berlin, Heidelberg, 2004), ECML’04, Springer-Verlag, pp. 217–226. doi:10.1007/978-3-540-30115-8\_22. 8

- [Lea51] LEAVITT H. J.: Some Effects of Certain Communication Patterns on Group Performance. *The Journal of Abnormal and Social Psychology* 46, 1 (1951), 38. 2
- [LZ15] LUO Q., ZHONG D.: Using Social Network Analysis to Explain Communication Characteristics of Travel-related Electronic Word-of-Mouth on Social Networking Sites. *Tourism Management* 46 (2015), 274–282. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0261517714001393>, doi:10.1016/j.tourman.2014.07.007. 3
- [Mes09] MESCH G. S.: Social Context and Communication Channels Choice Among Adolescents. *Computers in Human Behavior* 25, 1 (2009), 244–251. doi:10.1016/j.chb.2008.09.007. 2
- [MMKN08] MCLACHLAN P., MUNZNER T., KOUTSOFIOS E., NORTH S.: LiveRAC: Interactive Visual Exploration of System Management Time-Series Data. In *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08* (New York, New York, USA, 2008), Czerwinski M., Lund A., Tan D., (Eds.), ACM Press, p. 1483. doi:10.1145/1357054.1357286. 3
- [Mor34] MORENO J. L.: *Who Shall Survive? A New Approach to the Problem of Human Interrelations*, vol. 58 of *Nervous and mental disease monograph series*. Nervous and Mental Disease Publishing Co, Washington, D.C., 1934. 2
- [MQB19] MING Y., QU H., BERTINI E.: RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 342–352. doi:10.1109/tvcg.2018.2864812. 7
- [MS99] MANNING C. D., SCHÜTZER H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass. and London, 1999. 2, 3
- [Nui20] NUIX: Nuix Discover and Nuix Investigate, 2020. URL: <https://www.nuix.com/products>. 3
- [Pal20] PALANTIR: Gotham, 2020. URL: <https://www.palantir.com/palantir-gotham/>. 3
- [Pea11] PEARSON J. C.: *Human Communication*, 4th ed. ed. McGraw-Hill, New York, 2011. 2
- [PL08] PANG B., LEE L.: Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2, 1-2 (2008), 1–135. doi:10.1561/1500000011. 3
- [Rei10] REITZ F.: A Framework for an Ego-centered and Time-aware Visualization of Relations in Arbitrary Data Repositories. *arXiv abs/1009.5183* (2010). 3, 4
- [RK80] ROGERS E. M., KINCAID D. L.: *Communication Networks: Toward a New Paradigm for Research*. Macmillan USA, 1980. 2
- [ŘS10] ŘEHŮŘEK R., SOJKA P.: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (Valletta, Malta, 2010), University of Malta, pp. 46–50. URL: <http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf>. 3
- [Sco09] SCOLARI C. A.: Mapping Conversations About New Media: The Theoretical Field of Digital Communication. *New Media & Society* 11, 6 (2009), 943–964. doi:10.1177/1461444809336513. 2
- [Sco17] SCOTT J.: *Social Network Analysis*, 4th ed. SAGE, Los Angeles, 2017. 2
- [SD60] SAVAGE I. R., DEUTSCH K. W.: A Statistical Model of the Gross Analysis of Transaction Flows. *Econometrica* 28, 3 (1960), 551–572. URL: <http://www.jstor.org/stable/1910131>. 2
- [SFS\*19] SEEBACHER D., FISCHER M. T., SEVASTJANOVA R., KEIM D. A., EL-ASSADY M.: Visual Analytics of Conversational Dynamics. In *EuroVis Workshop on Visual Analytics (EuroVA)* (Porto, Portugal, 2019), von Landesberger T., Turkay C., (Eds.), EuroVA, The Eurographics Association. doi:10.2312/eurova.20191130. 2, 4, 5, 6, 10
- [Sim08] SIMMEL G.: *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Duncker & Humblot, Leipzig, 1908. 2
- [SK16] SALIMANS T., KINGMA D. P.: Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Advances in Neural Information Processing Systems* (2016), pp. 901–909. 7
- [SSG12] SHIRAVI H., SHIRAVI A., GHORBANI A. A.: A Survey of Visualization Systems for Network Security. *Transactions on Visualization and Computer Graphics* 18, 8 (2012), 1313–1329. URL: <http://ieeexplore.ieee.org/document/6007132/>, doi:10.1109/TVCG.2011.144. 3
- [Tri08] TRIER M.: Research Note —Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks. *Information Systems Research* 19, 3 (2008), 335–350. doi:10.1287/isre.1080.0191. 2
- [vH03] VAN HAM F.: Using Multilevel Call Matrices in Large Software Projects. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714)* (19-21 Oct. 2003), IEEE, pp. 227–232. doi:10.1109/INFVIS.2003.1249030. 3
- [WBJ74] WATZLAWICK P., BEAVIN J. H., JACKSON D. D. J.: *Menschliche Kommunikation: Formen, Störungen, Paradoxien*, 4th ed. ed. Hans Huber, Stuttgart, 1974. 2
- [WLY\*14] WU Y., LIU S., YAN K., LIU M., WU F.: OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media. *IEEE Transactions on Visualization and Computer Graphics* (2014). doi:10.1109/TVCG.2014.2346920. 3
- [XKS13] XIE J., KELLEY S., SZYMANSKI B. K.: Overlapping Community Detection in Networks: the State of the Art and Comparative Study. *ACM Computing Surveys* 45, 4 (2013), 1–35. URL: <http://arxiv.org/pdf/1110.5813v4>, doi:10.1145/2501654.2501657. 3
- [YEL10] YI J. S., ELMQVIST N., LEE S.: TimeMatrix: Analyzing Temporal Social Networks Using Interactive Matrix-Based Visualizations. *International Journal of Human-Computer Interaction* 26, 11-12 (2010), 1031–1051. doi:10.1080/10447318.2010.516722. 3
- [YP04] YOON B., PARK Y.: A Text-Mining-Based Patent Network: Analytical Tool for High-Technology Trend. *The Journal of High Technology Management Research* 15, 1 (2004), 37–50. doi:10.1016/j.hitech.2003.09.003. 2, 3
- [ZCCB12] ZHAO J., CHEVALIER F., COLLINS C., BALAKRISHNAN R.: Facilitating Discourse Analysis with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2639–2648. doi:10.1109/TVCG.2012.226. 3
- [ZLD\*15] ZHAO J., LIU Z., DONTCHEVA M., HERTZMANN A., WILSON A.: MatrixWave. In *Proceedings of the 33rd CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), Begole B., Kim J., Inkpen K., Woo W., (Eds.), CHI, ACM, pp. 259–268. doi:10.1145/2702123.2702419. 3