Topic-aware Network Visualisation to Explore Large Email Corpora

Tim Repke
Hasso Plattner Institute
Potsdam, Germany
tim.repke@hpi.de

Ralf Krestel
Hasso Plattner Institute
Potsdam, Germany
ralf.krestel@hpi.de

ABSTRACT

Nowadays, more and more large datasets exhibit an intrinsic graph structure. While there exist special graph databases to handle ever increasing amounts of nodes and edges, visualising this data becomes infeasible quickly with growing data. In addition, looking at its structure is not sufficient to get an overview of a graph dataset. Indeed, visualising additional information about nodes or edges without cluttering the screen is essential. In this paper, we propose an interactive visualisation for social networks that positions individuals (nodes) on a two-dimensional canvas such that communities defined by social links (edges) are easily recognisable. Furthermore, we visualise topical relatedness between individuals by analysing information about social links, in our case email communication. To this end, we utilise document embeddings, which project the content of an email message into a high dimensional semantic space and graph embeddings, which project nodes in a network graph into a latent space reflecting their relatedness.

1 INTRODUCTION

In our modern information society we produce substantial amounts of data each day. A large portion of it comes from the communication on social media platforms or through emails. Special graph databases enable the efficient storage of these large communication networks and provide interfaces to query or analyse the data. Visualising networks in their entirety on the other hand is a very challenging task. Users investigating a communication network want to find information about *when* does *who* communicate *with whom* about *what*. These kind of networks can be found in many different shapes. Modern social networks, such as Twitter or Facebook exhibit similar structures as classic, offline social networks [20]. We investigate another type of social network: a collection of emails.

Given the communication data over a year or more, it is practically impossible to gain an overview or quick insights into the latent network structure with a basic approach as shown in 1. Also, in such a traditional *network visualisation*, information about the content of messages sent between individuals is lost. Besides these traditional systems, more exotic approaches use the metaphor of geographical maps [17] to visualise networks, for example using topology to reflect connectivity of densely connected social communities. The map analogy can also be used to visualise the contents of documents by embedding them into a high dimensional semantic space [15] and projecting it on the map as a *document landscape*. In order to highlight how relationships form and change based on the interactions, the metaphor of a growing tree ca be used (ContactTrees [18]). Although this reflects temporal aspects of dynamic networks well, it focuses on one person as the root, thus an overview of the entire network is lost. CactusTrees [6] on the other hand represent hierarchical structures with the goal of untangling overlaid bundles of intersecting edges, making distant connections more apparent. As higher order dependencies may get lost in traditional visualisations, HoNVis [21] adds nodes to encode dependencies in chains of interactions. Usually, a communication network has many nodes and overlapping connections already, so Yang et al. [23] rather focus on discovering overlapping cores to improve the identification of community boundaries to highlight global latent structures. Similarly, Gronemann et al. [11] use the metaphor of islands and hills to visualise clustered graphs, making densely connected communities clearly noticeable. The edges are bundled and follow valleys of the resulting topology, thus making relationships between other communities hard to follow. MapSets [7] assume a graph that was laid out using embeddings reflecting communities. An algorithm then draws regions around clusters of nodes, such that the bounding shapes are contiguous and non-overlapping, but yet abstract. Another approach to visualise networks at full scale is to aggregate nodes based on their spatial distribution and thereby allowing for a simple exploration with contour lines and heatmap overlays to emphasise latent structures as proposed by Hildenbrand et al. [13].

Document visualisation aims to visualise the content, such that users gain quick insights into topics, latent phrases, or trends. Tiara [22] extracts topics and derives time-sensitive keywords to depict evolving subjects over time as stacked plots. Other approaches project documents into a latent space, using topic models or embeddings. Creating scatter-plots of embedded documents of a large corpus may result in a very dense and unclear layout, so Chen et al. [4] developed an algorithm to reduce overfull visualisations by picking representative documents. A different approach is taken by Fortuna et al. [8], who do not show documents directly, but generate a heatmap of the populated canvas and overlay it with salient phrases at more densely populated areas from the underlying documents in that region. Friedl et al. [10] extend that concept by drawing clear lines between regions and colouring them. They also add edges between salient phrases based on co-occurrences in the texts. Most recently Cartograph [19] was proposed, which is visually very similar to previous approaches, but uses pre-rendered information of different resolution and map technology to enable a responsive interactive visualisation. Regions are coloured based on underlying ontologies from a knowledge-base.

Our goal is to *merge approaches for network and document visualisations* in one interactive user interface. This means to integrate multiple dimensions of email datasets including time, interactions, users and topics into a 2D map representation. Giving an overview over latent structures and topics in one map may significantly improve the exploration of a corpus by users

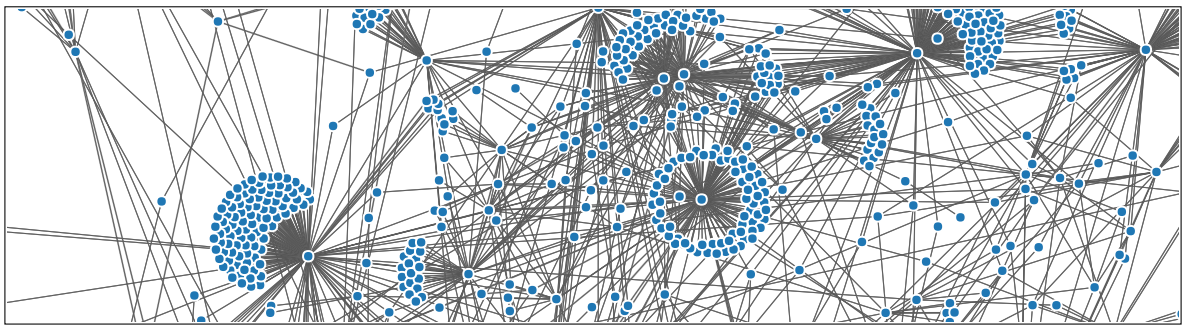

Figure 1: Traditional basic visualisation of a communication graph from 2000 emails using force layout

unfamiliar with the domain and terminology. Also domain experts could benefit from such an overview, e.g. by easily being able to identify global patterns in the data.

A specific application scenario that could benefit from such integrated, interactive visualisations is the analysis of large, unstructured, heterogeneous data collections. Data-driven journalism [5] often has to deal with leaked, unstructured, very heterogeneous data, e.g. in the context of the Panama Papers, where journalists needed to untangle and order huge amounts of information, search entities, and visualise found patterns [3]. Similar datasets are of interest in the context of computational forensics [9]. Auditing firms and law enforcement need to sift through huge amounts of data to gather evidence of criminal activity, often involving communication networks and documents [14].

2 INTERACTIVE VISUALISATION

Systems for document exploration largely vary in what they display and how users interact with them. This depends partly on the available raw data, but also on information extracted from pre-processing or enrichment with external sources. 1 shows a basic visualisation of the network graph extracted from an email corpus. Although it is an improvement over only listing connections, large densely connected graphs quickly become hard to read and information about the email contents is lost.

Exploring document collections can be seen as a top-down approach, where the system provides abstract overviews of the entire document collection and users incrementally refine the search, narrowing the results to just a few documents of interest. Such a top-down approach may help users without prior knowledge to get a sense for the data by visualising high level latent structures of communication networks or the topical distributions.

In the scope of this work we primarily consider documents to be emails or data attached to them. The sender, recipients, time, and content can directly be extracted from the raw data. We call these – and results from further processing – *dimensions* that can be visualised. From the contents one may infer named entities, topics, embeddings, or salient phrases, while the communication network spanned by sender-recipient pairs can be used to detect salient structures and hierarchies. The temporal information enables the previously mentioned data to be analysed over time to detect evolving or changing patterns.

There are numerous ways to visualise each dimension on its own or in combination with others. The requirement of a dimension and its priority in a visualisation is dictated by the system objective. From the wide range of possibilities, we strive for a system which supports the exploration of a large collection of documents without any prior knowledge about its content and individuals involved.

In our system, we use the names and email addresses of senders and recipients (*individuals*), communication network, semantic vector representations of email contents, and as part of an overlay the timestamps of emails and propose a graph layout over a document landscape that visually describes *who* talks *with whom* about *what* at a given *time period*.

3 SYSTEM ARCHITECTURE

Visualising communication networks in a topic-aware fashion to explore documents and salient structures is not straightforward Different layout objectives may produce contradicting results and the challenges of processing big data need to be addressed [2]. In this section, we describe algorithmic approaches behind the system we are working on. For a discussion of engineering aspects on how to store, serve, and render the map-like data, we refer to the Cartograph stack [19], as we will focus on the process how to get the information that the map is generated from.

We visualise the embedded emails as dots in a two-dimensional landscape in which individuals are placed as nodes connected by edges. All emails between two individuals are reduced into one edge reducing visual complexity and making it easier to detect salient structures. However, that comes with the trade-off that nodes and edges cannot be perfectly placed in the landscape to cover all semantic aspects of the communication between them, but rather an estimate. Our very early prototype placed some individuals with no dominant topic in a crowded area in the centre of the landscape as shown in 2, where colours of opaque dots for emails correspond to that of the sender. Although the network visualisation at this point does not make connections more clear than in 1, users can already distinguish individuals with similar or unrelated topics.

Our proposed algorithm to find a stable network layout has three stages, namely an *(i) initialisation phase* which creates the landscape and roughly places nodes and connections, an *(ii) update phase* which iteratively updates the node placement towards a better fit, and finally a *(iii) post processing phase* where edges become splines to make latent structures more clear and a map topology is added.

Initialising the Landscape. To generate the document landscape, we first process the network graph to roughly determine regions, where documents will be placed. Therefore we apply node2vec [12] to the communication network and embed each individual's node. We separate the graph into communities $P_i \subset P$ using the kernel density of the resulting populated space at

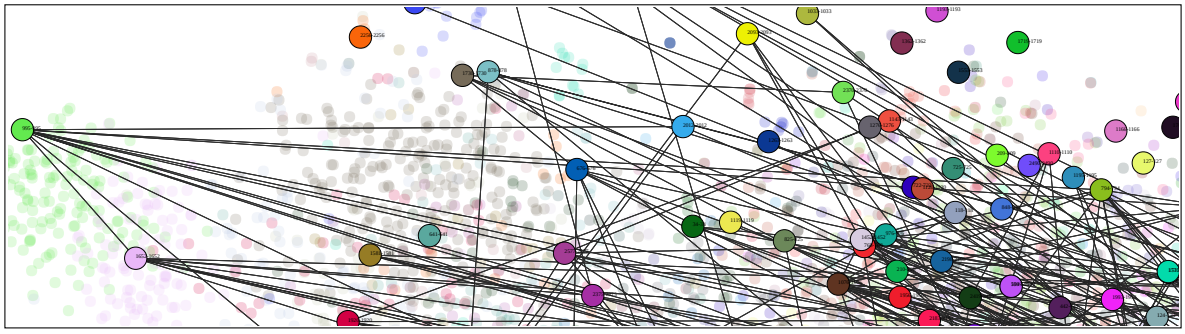

Figure 2: Rendered prototype output after landscape initialisation without prior community segmentation

threshold κ, where a higher κ results in more, but smaller communities. For each community P_i, pairwise neighbourhood similarities are calculated using euclidean distance between nodes, forming the triangular matrix S_i, where s_{kl} is the similarity between $p_k, p_l \in P_i$.

Furthermore, we train document embeddings [15, 25] on all emails and use them to infer high dimensional semantic vector representations. Let M_i be the set of emails that originated in community P_i. For each email $m \in M_i$, the dimensionality is reduced using t-SNE [16], which retains possible semantic clusterings of documents in the higher dimensional space. The resulting two-dimensional vectors are then placed as dots on the map using the centre of embedded network communities as the respective origin, whereas the size is determined by the number of related individuals.

We also initialise communication network's layout. Thereby, the staring position of a node representing an individual is determined by the normalised sum of two-dimensional vectors of all emails he or she has sent or received. This way, we implicitly group semantically related individuals into communities as frequent communication biases this normalised sum. Straight edges are added between the nodes if the respective individuals exchanged emails. Note, that many edges may only represent a small number of emails. Applying a variable threshold σ can reduce the computational load in later stages, as these edges will not impact the overall layout very much. They can be added again as the user requests a detailed visualisation by zooming in or through other interactions.

In the algorithm's second stage, we iteratively try to improve the layout of the communication network by finding a balance between the closeness of nodes to semantic context and densely connected neighbourhoods a node belongs to. Therefore, for each individual $p_j \in P$ we use linear regression to fit a line $\widehat{m_{p_j}}$ though all two-dimensional vectors of emails he or she has sent or received. As a node is placed near this line, it remains in a semantically good position.

Adjusting the Network Layout. The first stage of our proposed algorithm produces a fixed document landscape and roughly fits the communication network on top. We now aim to incrementally adapt the layout of the graph to better reflect salient structures in the network while keeping each individual's node close to the reflective semantic area in the landscape.

Therefore we define a score quantifying how well the current layout fits these objectives:

$$\sum_{p_i \in P} \left[\eta \, d(p_i, \widehat{m_{p_i}}) + \sum_{p_j \in P} \theta \left(s_{ij} - d(p_i, p_j) \right) \right] \tag{1}$$

where $d(\cdot, \cdot)$ is the distance between two nodes (zero if no connection exists) or shortest distance from a node to its ideal line. To adjust the layout towards either a better semantic or structural fit, we introduce parameters θ and η.

In order to minimise 1, we use stochastic gradient descent. In each iteration step, we can derive the direction and magnitude each node should be moved towards a better semantic fit and closer proximity to it's neighbourhood in the network.

The semantic gradient for $p_j \in P$ is defined by

$$\vec{\delta_j^s} := (p_j^m - p_j) \, \|p_j^m - p_j\| \tag{2}$$

where p_j^m is the closest point on $\widehat{m_{p_j}}$ to p_j and $\|\cdot\|$ denotes the euclidean norm, while neighbourhood gradient is defined by

$$\vec{\delta_j^n} := \sum_{\substack{p_k \in P \\ p_j \sim p_k}} (p_j - p_k) \, \left(\|p_j - p_k\| - s_{jk} \right) \tag{3}$$

where $p_j \sim p_k$ denotes that an edge exists between p_j and p_k.

With the definitions in 2 and 3, we can formulate the update vector $\vec{\delta_j}$ for node $p_j \in P$ as

$$\vec{\delta_j} := \xi \left(\theta \vec{\delta_j^n} + \eta \vec{\delta_j^s} \right) \tag{4}$$

where ξ is the learning rate and θ, η as before parameters to weight between a better semantic or neighbourhood fit.

Most likely, complex network structures might prevent the stochastic gradient descent to find a stable minimum, so the score of the objective function should be monitored or intermediate layouts be visually evaluated to determine a satisfactory result.

Post Processing. Lastly, we use the post processing stage to enhance the readability of our visualisation. Densely connected communities in the graph are potentially hard to read, thus we apply edge bundling [1] to visually clear latent structures. We also apply MapSets [7] to separate the regions for each community. Since semantically similar emails may appear in different communities, we apply colouring based on clusters in the original global document embedding space to retain this aspect. Choosing the colours depends on the number of latent topics that should be depicted [24]. If the topic number exceeds 25-30 topics, grouping topics and allowing for zooming within a two-level topic-hierarchy ensures distinguishable colors for up to 10 subtopics ($25 \times 10 = 25$). In order to represent temporal aspects of the data, we calculate the kernel density of the document landscape for fixed time-intervals, which can be used to add heat-map overlays that users can select later on.

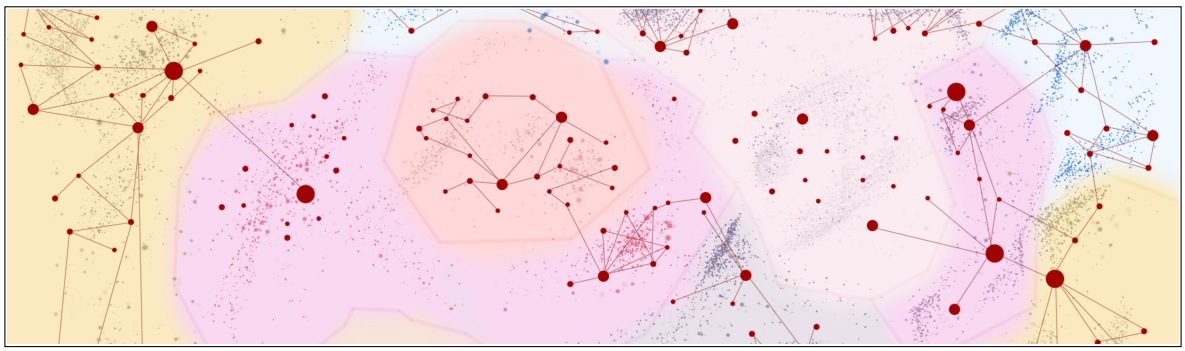

Figure 3: Semantic landscape of email contents and dominant communication patterns (drawn mockup)

4 CONCLUSION AND VISION

In this paper, we described an algorithm to lay out a communication network on top of a landscape of semantically embedded emails. This is still work in progress, thus 3 shows only a manually drawn mock-up of the visualisation we envision. In it, individuals are represented as nodes positioned such that densely connected communities are visually clustered. Edges describe the email traffic, where the opacity and thickness is used to indicate the frequency of messages between the nodes they connect.

The semantic representations of emails are used to place dots on a background layer which we call the *document landscape*. This landscape is used as additional input to the graph layout algorithm, aiming to place a node within corresponding semantic regions. The colouring of regions in the landscape is derived from densely connected communities in the communication graph. Optionally, representative words are selected for densely populated areas in the landscape, so that users get a rough idea about subjects in that area. The aforementioned timestamps of emails can be used to generate a heatmap overlay to show the activity in a certain time interval which is controlled by a slider. Similar to modern geographical maps, zooming into a region reveals more details. In our case, less prominent individuals and their connections are shown along with additional salient phrases from the document landscape. Selecting a node will not only highlight connected edges but may also temporarily show more edges which were previously hidden at that zoom level. The user will also be able to retrieve documents with the help of a selection rectangle or clicking dots in the document landscape.

In future work, we hope to evaluate this system using full-scale real-world data as well as practitioners from journalism and auditing. It may also be interesting to experiment with embedding methods, which take both the emails and the network graph as input and directly project the inferred representations into the two-dimensional landscape to simplify the proposed algorithm.

REFERENCES

[1] Benjamin Bach, Nathalie Henry Riche, Christophe Hurter, Kim Marriott, and Tim Dwyer. 2017. Towards unambiguous edge bundling: Investigating confluent drawings for network visualization. *Transactions on Visualization and Computer Graphics* 23, 1 (2017), 541–550.

[2] Nikos Bikakis and Timos Sellis. 2016. Exploration and visualization in the web of big linked data: A survey of the state of the art. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference*, Vol. 1558. CEUR-WS.org.

[3] Marie-Anne Chabin. 2017. Panama papers: a case study for records management? *Brazilian Journal of Information Science: Research Trends* 11, 4 (2017).

[4] Yanhua Chen, Lijun Wang, Ming Dong, and Jing Hua. 2009. Exemplar-based visualization of large document corpus. *Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1161–1168.

[5] Mark Coddington. 2015. Clarifying journalismâĂŹs quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism* 3, 3 (2015), 331–348.

[6] Tommy Dang and Angus Forbes. 2017. CactusTree: A tree drawing approach for hierarchical edge bundling. In *Proc. of the Pacific Visualization Symposium*. IEEE, 210–214.

[7] Alon Efrat, Yifan Hu, Stephen G Kobourov, and Sergey Pupyrev. 2015. MapSets: Visualizing Embedded and Clustered Graphs. *Journal of Graph Algorithms and Applications* 19, 2 (2015), 571–593.

[8] Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. 2005. Visualization of text document corpus. *Informatica* 29, 4 (2005), 497–502.

[9] Katrin Franke and Sargur N Srihari. 2007. Computational forensics: Towards hybrid-intelligent crime investigation. In *International Symposium on Information Assurance and Security*. IEEE, 383–386.

[10] Daniel Fried and Stephen G Kobourov. 2014. Maps of computer science. In *Proc. of the Pacific Visualization Symposium*. IEEE, 113–120.

[11] Martin Gronemann and Michael Jünger. 2012. Drawing clustered graphs as topographic maps. In *Proc. of the Symposium on Graph Drawing and Network Visualization*. Springer, 426–438.

[12] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proc. of the Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[13] Jan Hildenbrand, Arlind Nocaj, and Ulrik Brandes. 2016. Flexible Level-of-Detail Rendering for Large Graphs. In *Proc. of the Symposium on Graph Drawing and Network Visualization*. Springer, 625–627.

[14] Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan. 2008. An intelligent system for semantic information retrieval information from textual web documents. In *International Workshop on Computational Forensics*. Springer, 135–146.

[15] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of the International Conference on Machine Learning*. PMLR, 1188–1196.

[16] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[17] Patrick Cheong-Iao Pang, Robert P Biuk-Aghai, Muye Yang, and Bin Pang. 2017. Creating realistic map-like visualisations: Results from user studies. *Journal of Visual Languages and Computing* 43 (2017), 60–70.

[18] Arnaud Sallaberry, Yang-chih Fu, Hwai-Chung Ho, and Kwan-Liu Ma. 2016. Contact trees: Network visualization beyond nodes and edges. *PLOS ONE* 11, 1 (2016), 1–23.

[19] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking Spatial Visualization Through Semantic Enhancement. In *Proc. of Conference on Intelligence User Interfaces*. ACM, 179–190.

[20] Kaveri Subrahmanyam, Stephanie M Reich, Natalia Waechter, and Guadalupe Espinoza. 2008. Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology* 29, 6 (2008), 420–433.

[21] Jun Tao, Jian Xu, Chaoli Wang, and Nitesh V Chawla. 2017. HoNVis: Visualizing and Exploring Higher-Order Networks. In *Proc. of the Pacific Visualization Symposium*. IEEE, 1–10.

[22] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. TIARA: a visual exploratory text analytic system. In *Proc. of the Conference on Knowledge Discovery and Data Mining*. ACM, 153–162.

[23] Jaewon Yang and Jure Leskovec. 2014. Overlapping communities explain core–periphery publisher of networks. *Proc. IEEE* 102, 12 (2014), 1892–1902.

[24] Achim Zeileis, Kurt Hornik, and Paul Murrell. 2009. Escaping RGBland: selecting colors for statistical graphics. *Computational Statistics & Data Analysis* 53, 9 (2009), 3259–3270.

[25] Zhaocheng Zhu and Junfeng Hu. 2017. Context Aware Document Embedding. *CoRR* abs/1707.01521 (2017).