Erschienen in: Computer Graphics Forum ; 32 (2013), 3pt4. - S. 371-380
https://dx.doi.org/10.1111/cgf.12124

Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature

D. Oelke[1], D. Kokkinakis[2] and D. A. Keim[3]

[1]Ubiquitous Knowledge Processing Lab (UKP), DIPF, Frankfurt, Germany
[2]Språkbanken Group of the Department of Swedish, University of Gothenburg, Sweden
[3]Data Analysis and Visualization Group, University of Konstanz, Germany

Abstract

In prose literature often complex dynamics of interpersonal relationships can be observed between the different characters. Traditionally, node-link diagrams are used to depict the social network of a novel. However, static graphs can only visualize the overall social network structure but not the development of the networks over the course of the story, while dynamic graphs have the serious problem that there are many sudden changes between different portions of the overall social network. In this paper we explore means to show the relationships between the characters of a plot and at the same time their development over the course of a novel. Based on a careful exploration of the design space, we suggest a new visualization technique called Fingerprint Matrices. A case study exemplifies the usage of Fingerprint Matrices and shows that they are an effective means to analyze prose literature with respect to the development of relationships between the different characters.

1. Motivation

Literature can be studied in a number of different ways and from many perspectives, but text analysis will surely always make up a central component of literature studies. Our work aims at supporting literature scholars in this task by providing them with visualizations that make patterns or trends with respect to a certain text property easy to perceive. Specifically, the approach presented in this paper concentrates on the development of social networks in prose literature that represent the relationships between characters in a novel. The visualization of such networks can reveal inherent patterns like subgroups of characters that interact with each other. However, usually the relationships in a novel are not static but develop during the plot. Social networks build up gradually and some acquaintances may only be important for part of the story. The goal of our work is to enable literature scientists to dig deeper and explore a novel in terms of where in the plot certain protagonists are related to each other. This way a deeper understanding of the structure of the novel with respect to co-occurrences of characters becomes possible and more details of the story line are revealed. The basic idea of the paper is to visualize pairwise relations between characters in so-called co-occurrence glyphs that can be considered a fingerprint of the dynamics between two protagonists. The fingerprints are arranged in an adjacency matrix to get the overall picture of the storyline.

The paper is structured as follows: First, some background information about the applied natural language processing techniques as well as related work is given in Section 2. This is followed by a careful consideration of the design space in Section 3 which motivates the visualization technique that is introduced in Section 4. Section 5 explains how to read Fingerprint Matrices. This is further exemplified in the case studies in Section 6 in which a modern English novel and a Swedish novel of the 19th century are analyzed. The paper concludes with a summary and future work.

2. Background

2.1. Visual literature analysis

Traditionally, literature scholars analyze a novel by reading it sequentially. In contrast to this, the literature scholar Franco Moretti coined the idea of "distant reading" and suggests to "reduce the text to a few elements, and abstract them

from the narrative flow, and construct a new, artificial object" [Mor05]. In his work he makes use of common diagrams like line charts, maps, or evolutionary trees to visualize and analyze meta data about a book collection or certain aspects of a text. Vuillemot et al. [VCPK09] employ word clouds and self-organizing graphs for literature analysis and present a tool that permits to analyze a novel interactively with respect to several properties. In [PRY*06] a tabular representation that is enriched with visual symbols is used to present the results of an automatic algorithm for detecting erotic statements. Rydberg-Cox [RC11] and Oelke et al. [OKM12] both generate social network graphs of characters in prose literature. [RC11] additionally employs scatterplot views that allow the user to search for correlations between several variables of the meta data that come with the novels. [OKM12] complements the node link diagrams with summary plots and pixel-based visualizations. Rohrer et al. [RES98] experimented with using implicit surfaces to compare single documents with respect to the most frequent terms and to visualize a document collection. Kim et al. [KKEE11] introduce WordBridge, a graph-based visualization techniques that depicts relations between entities by means of tag clouds.

Also related to our work are text analysis techniques that do not focus on prose literature but consider temporal developments (i.e., dynamics). However, in contrast to our work most approaches such as [CLT*11, WLS*10, CVW09] consider dynamics *between* different documents instead of the developments of a storyline within a single document. (Techniques that consider dynamics within a document are reviewed in Sections 2.2 and 3.2.)

2.2. Dynamic social network visualization

Most approaches for visualizing dynamic social networks use animation to connect a series of node-link diagrams (cf. [MMBd05, BW97]). Animation seems like a natural encoding of the temporal dimension but it requires to keep the node movement to a minimum to preserve the mental map of the user between different steps of the animation [PHG07]. Alternatively also small multiples of graphs (one for each time step) can be displayed.

Other approaches for visualizing the dynamics in social networks include [BC02] that stacks several semi-transparent planes on top of each other whereby each layer shows a modification of the network in the course of time. Reda et al. [RTJ*11] as well as Tanahashi and Ma [TM12] represent the temporal dimension along the x-axis. The y-axis in the visualization is used to separate the different previously determined communities (subgroups) from each other. Entities are represented as threads whose path shows the community affiliation of a certain person over time. Burch et al. [BVB*11] suggest to change the layout of a node-link diagram in a way that nodes are arranged on a vertical line with two such lines representing one graph. The nodes in the two stripes are connected if an edge exists between them in the network. Time is represented by the horizontal axis along which the sequence of graphs is displayed.

Most related to our work are approaches that encode the relationships between the entities of a social network in an adjacency matrix. Yi et al. [YEL10] represent the graph in a so called TimeMatrix, an adjacency matrix whose cells encode the temporal dimension of a certain attribute using glyph-based visualizations (e.g., a bar chart). Because the glyphs are not visible anymore if a large graph is displayed, the authors propose the usage of semantic zooming. Furthermore, neighboring cells can be aggregated and range sliders permit to filter out specific nodes or edges. In [BN11] glyphs that are based on the two visual variables angle and length are used to enable an exploration of evolving dyadic relations. The work that is most similar to our approach is [SWS10] in which also pixel-based glyphs are used to show the dynamics of social networks in an adjacency matrix. In contrast to this approach, in our visualization a pixel does not represent one step in time but a text unit of a document (e.g., a sentence). While it would be possible to generate one graph per position of the pixel-based glyphs proposed in [SWS10], we do not discretize the temporal dimension (in our case the text sequence) but represent a relation between two entities with a sequence of pixels. Consequently, relations are defined by local patterns (which could be seen as a time span) rather than by single pixels (points in time).

2.3. Relationship detection in novels

In the literature analysis task that is addressed in this paper, we are particularly interested in the relations between certain characters. Relation extraction is the task of recognizing and characterizing semantic relations among a pair of e.g., concepts or known named entities in text. Relation extraction approaches can be classified in various ways. Knowledge engineering approaches (e.g., rule-based, linguistic based), learning approaches (e.g., statistical, machine learning, bootstrapping) and hybrid ones; for a general review of relation extraction techniques see [Hac09].

What is aimed for in our scenario, which is distinctive from previous approaches, is that we do not only want to learn *which characters are related* to each other in some part of the story, but we also need to know *where* in the plot they live their relationship. This means that we need a measure that does not only focus on explicit declarations of relationships (e.g., X is the mother of Y), but enables us to detect subsections that report on established relationships. Therefore, we employ a co-occurrence-based measure, which is based on the assumption that a high density of co-occurrences in some part of the plot permits the conclusion that the two characters are related. Note that in the current work the relation extraction is based solely on explicit mentions of the names of the characters (disregarding references with personal pronouns etc.).

3. Design Considerations

In prose literature often complex dynamics of interpersonal relationships exist between different characters. Traditionally, static node-link diagrams are used to depict the global social network of a novel. While the overall structure of the social network can be well perceived in such a diagram, some questions remain unanswered by a static, global relationship graph. These include:

- When is a certain character introduced to the plot?
- When do certain characters meet the first time?
- In which order are the acquaintances made?
- How intense is a relationship in some part of the novel?
- What subgroups of characters exist and how do they evolve during the plot?

More generally speaking we could ask: How does the social network of characters evolve in the course of the novel?

3.1. Node-link diagrams

Often animated node-link diagrams are used to visualize dynamic social networks. To find out if animated graphs (or a series of static node-link diagrams) could be a suitable means to visualize the development of social networks in prose literature, we not only generated a sample animated graph for one of the novels in our case studies (see Section 6.1 and supplementary material) but also analyzed a set of 10 Swedish novels of the 19th/20th century with respect to the properties of their evolving network. Character names in the novels were identified with the help of a named entity recognition system that was adapted to the language used around the turn of the 20th century [BK10]. Using a sliding window approach, for each block of 100 sentences a network structure was generated. Neighboring blocks overlap each other by 10 sentences to reduce the danger of introducing artifacts that may be caused by arbitrarily chosen split points. An edge is inserted between two characters if their names appear at least once in the block in the same sentence.

Discussion of the results:

- *Size of the networks and number of time steps*
  On average, the graphs had 13.8 edges (min = 0, max = 88) and 7 nodes (min = 0, max = 20). Thus, most networks are rather small and can be displayed with a node-link diagram. The number of time steps in the animation depends on the aggregation level that the networks are calculated for. With the chosen window size, the average number of blocks in our experiment was 64.5 (min = 22, max = 87). Note that the higher the aggregation level is, the more details are missed because the relationships in a novel are not static but develop and change constantly in the course of the novel. On the other hand, a lower resolution results in more networks (and thus, longer animations or quicker changes) and hampers the detection of higher-level structures.
- *Number of changes between subsequent networks*
  If there are many changes from one image to the next in an animation, it becomes difficult for the user to follow the development of the network structure. A recently published paper [FQ11] reports on results of a user study in which animated graphs are compared to static graph sequences. Interestingly, there was clear evidence in favor for static images. Throughout most experiments the users were faster and sometimes more accurate when using static graph sequences instead of the animation. The authors hypothesized that this was due to the fact that between single time steps many nodes were moving which distracted the users. With an average of 16.9 edge changes (min = 0, max = 84), i.e. an edge being removed or added, and 4.9 node changes (min = 0, max = 17) from one step to the next, this finding is relevant for our task. This is aggravated by the fact that the number of different characters per book (= unique nodes in the graph sequence) is quite high (average = 22.2, min = 9, max = 39, excluding characters that are mentioned less than 10 times in a novel).
- *Influence of parameter settings*
  For generating the graph sequences, several critical choices in terms of parameter settings had to be made: Besides choosing the right aggregation level as discussed above, we also had to decide on a measure for detecting relationships. Setting the co-occurrence threshold to 0 (names have to be in the same sentence), certainly will miss some links, but also avoids false positives. Alternatives include: considering co-occurrences of names that are separated by several sentences or requiring several co-occurrences in a block. All of these definitions have their strengths and weaknesses but are somewhat arbitrary. However, representing the data in a node-link diagram forces us to take a decision to decide if an edge should be inserted between two characters or not.

Based on the above observations, we conclude that using animated node-link diagrams for the task is not as straightforward as it may first seem. When applied to the domain of literature analysis, animated graphs come with several disadvantages that we believe can be avoided with a technique that is better adapted to the special characteristics of the dynamic social networks in prose literature.

3.2. Adjacency matrices

Another common graph representation are adjacency matrices. In contrast to node-link diagrams, adjacency matrices come with the advantage that edge crossings and overlapping nodes are not a problem and thus the technique can also deal with large and dense graphs. However, their successful application is highly dependent on a meaningful ordering of lines and columns that reveals inherent clusters and subgroups and they are sometimes considered less intuitive than node-link diagrams.

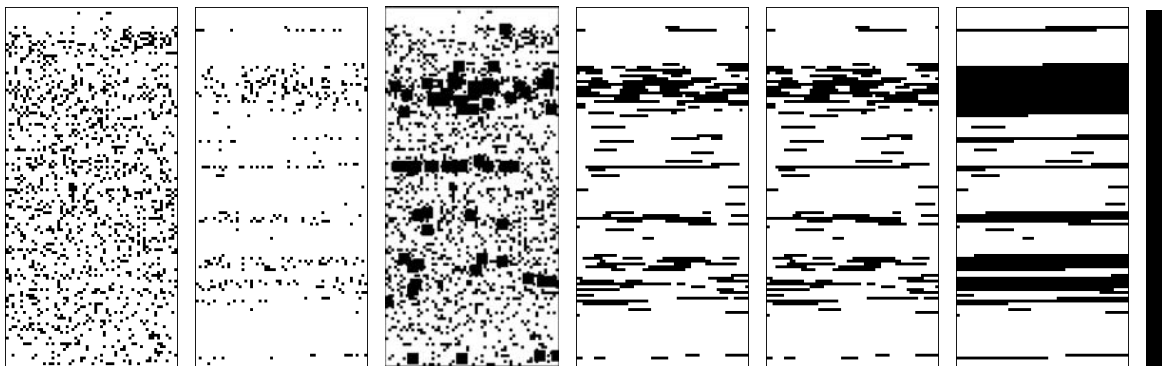


Figure 1: *Motivation of the design of the fingerprint glyphs. (a) Fingerprint in which sentences that mention the name Harry are highlighted. (b) Fingerprint in which mentions of the name Hagrid are highlighted. (c) Mentions of Harry and/or Hagrid are shown in the same plot (green = both in same sentence). (d) Co-occurrence lines instead of single mentions (see also Figure 2) (e) Co-occurrence lines with coloring as a redundant encoding of the position. (f) Colored co-occurrence lines for which the inherent block structure was boosted. (g) The colormap that is used to encode the position of a co-occurrence line.*

If adjacency matrices are used to visualize dynamic social networks, the temporal dimension has to be encoded with special glyphs in the cells of the matrix (cf. [YEL10, BN11, SWS10] that inspired our work). In our case there is no temporal dimension, but we would like to show the development of the social network in the course of the document.

Commonly, the distribution across a document is shown by representing the document as a rectangle that is split up into smaller units (e.g., with rectangles or stripes that each represent one unit of text) which are color-coded to show a text property [Hea95, BE96, DZG*07, KO07, FD00, RA00]. Alternatively, glyphs can be used to mark the position of a text property [XS07, FLE06]. Sometimes the terms are directly highlighted in the text, e.g., in a document thumbnail [KMOZ08] or are marked by adding a second graphical layer (text overlay) as in [WFR*01] or InkBlots [AC07].

4. Fingerprint Adjacency Matrix

In the following we motivate our design decisions and introduce fingerprint matrices as a technique for displaying dynamic social networks of prose literature. The basic idea is to represent the data in an adjacency matrix in which each row/column represents one character of the novel. The cells of the matrix are filled with pixel-based glyphs (fingerprints) that show the development of the observed co-occurrences across the document. The pixel-based representation of the document allows us to represent the data at a high resolution level and to refrain from setting arbitrary split points in the novel as would be necessary if we displayed the data in an animated node-link diagram. Furthermore, the fingerprints are designed in a way that the co-occurrences are directly visualized which provides the user with the necessary transparency to interpret the data correctly. This is important because co-occurrence-based measures do not take the semantics of the text into account which naturally implies that there is a risk for some false positive relations.

4.1. Fingerprint glyphs

The design of our co-occurrence glyphs is inspired by pixel-based techniques such as [FD00, KO07, BE96]. In the fingerprints each sentence is represented with a pixel and the pixels are aligned from left to right and top to bottom. In the fingerprints of Figures 1(a) and 1(b) color was used to highlight sentences that mention a certain character (in this case *Harry* or *Hagrid*). In this representation it is easy to see that *Harry*, the main protagonist of the novel, is almost always present whereas *Hagrid* just shows up from time to time.

Since we are interested in the co-occurrence of two characters, we integrate both figures into one. In Figure 1(c) again sentences in which *Harry* is mentioned are colored in blue and the ones containing the term *Hagrid* in pink. In addition sentences mentioning both characters are highlighted in green and are boosted with semi-transparent halos to make them visually more salient (see [OJS*11] for more details about boosting techniques for pixel-based visualizations).

The relationship extraction method that we use is based on co-occurrences, thereby assuming that the closer two characters are to each other, the more likely it is that they have a relationship. But how close do they have to be? Obviously, the closer two mentions are, the more likely it is that the two characters co-occur. However, it is difficult to determine a universally valid threshold because the narration of a dialog or an event may be interrupted to provide the reader with

more general information about the situation that the characters are in or anaphora (such as personal pronouns) are used in between to denote the characters. Consequently, our fingerprinting technique has to be designed in a way that the notion of co-occurrence is transparently shown to the user which implies that the choice of the threshold value becomes less critical as will be detailed below.

Another issue becomes obvious in Figure 1(c): when searching for pixels with two different colors that are close to each other in the sequence of the text, it is very difficult *not* to search for such proximity in the vertical direction as well. In a two-dimensional plot it seems natural to search for patterns in 2D. Thus, this kind of representation apparently introduces some visual artifacts and misleads the interpretation. The problem is alleviated if the fingerprints are produced as thin long rectangles instead of squares because this means that vertically neighboring pixels represent pixels that are not too distant in the course of the novel. To further overcome this disadvantage, we connect neighboring pixels of different color if their distance is below a user-specified co-occurrence threshold (see Figures 1(d) and 2). As a consequence, we leave the space of classical pixel-based visualizations and get more specialized towards visualizing co-occurrences. The usage of horizontal stripes comes with the advantage that the eye of the user is guided and at the same time the notion of defining relation as near-by occurrence of two characters is transparently shown. Moreover, irrelevant passages, in which no co-occurrence can be observed, are filtered out which further reduces noise and focalizes the analyst's attention to the main characteristics of the data.

To ease the comparison between different fingerprints, we use color to redundantly encode the position of the co-occurrence lines in the document (see Fig. 1(e), 1(g)). Finally, we visually boost the saliency of passages with a high density of co-occurrences by connecting two co-occurrence lines if their distance is below a user-specified threshold (see Fig. 2). In Figure 1(f) semi-transparent colors were used for boosting to permit a distinction of these pixels from the "real" co-occurrences. In the example shown we increase the transparency with the distance of the pixel from the two lines it connects. Alternatively, the transparency level of the complete line could be set in correlation to its overall length. Using such a boosting technique, the main structure of a fingerprint becomes easier to perceive (compare Fig. 1(e) to 1(f)).

4.2. Choosing appropriate parameters

There are two parameters that need to be specified by the user: the co-occurrence and the boosting threshold. The co-occurrence threshold specifies how many sentences apart the mentions of two characters may be to still consider them as co-occurring. In general it can be said that choosing higher values for this threshold increases the chance for false positives but also decreases the risk to miss relations. In our experience values between 3 and 6 have proven successful. Note that choosing an appropriate threshold value is less critical for fingerprints than for sequences of node-link diagrams. Typically, in a section in which two characters heavily interact, multiple co-occurrences are observed that form a visual block structure in the fingerprint. Therefore, a mistakenly recorded co-occurrence will either lead to a single outlier stripe or visually strengthen an otherwise correct block structure.

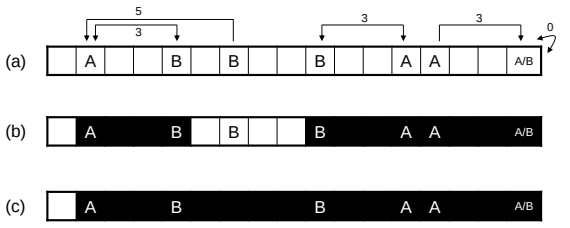


Figure 2: *Illustration of co-occurrence threshold and boosting. (a) Sentences marked with A,B mention the names of (one of) the two characters in focus. The numbered arrows show the distance, i.e. number of sentences, to the closest mention of the other character. (b) Given a co-occurrence threshold of 3, the highlighted sections would be colored. (c) Given a boosting threshold of 5, the two sections would be linked with semi-transparent coloring (transparency increases with the distance from the colored snippets).*

Boosting is used to make existing block structures visually more salient. The semi-transparent coloring of the boosted sequences allows the user to see where two co-occurrence lines were bridged which permits to verify and adjust the setting accordingly. In general, the boosting can be seen as a visual aggregation or smoothing which eases the perception of higher-level structures of the co-occurrence network. See supplementary material for a comparison of fingerprints with different boosting thresholds.

4.3. Fingerprint Matrix

For each pair of persons A and B a fingerprint glyph is generated. These glyphs are then arranged in an adjacency matrix (see Figure 3). Although the matrix is symmetric and therefore, the matrix entries $m_{A,B}$ and $m_{B,A}$ are the same, the full matrix is displayed because fingerprint matrices are usually read line-wise. We use the diagonal to show single person fingerprints like the ones shown in Figures 1(a) and 1(b) because knowing where the name of a person was mentioned can be important for interpreting the data (see Section 5). We slightly boost the pixels in the single person fingerprints with semi-transparent halos to increase their visual saliency.

In an adjacency matrix, ordering the rows and columns according to the intended analytic perspective is critical for

a successful application. The similarity function that we use counts the number of pixels that are colored in both fingerprints which groups fingerprints with a similar block structure. However, depending on the analytic task alternative orderings could be applied.

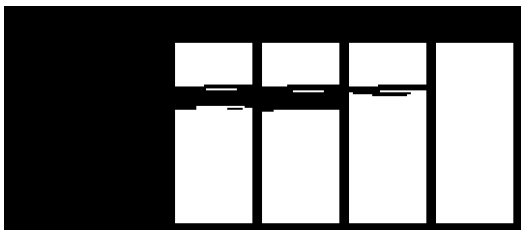
To increase the scalability of the technique (in terms of being able to display more characters or longer books), instead of sentences, paragraphs could be used as the lowest level of the analysis. Experiments have shown that a weak aggregation (e.g., 10 sentences per paragraph) still yields good results, whereas low resolution levels camouflage many details, the same as animated graphs do. (See supplementary material for an aggregated view of the data displayed in Figure 3.) Furthermore, aggregation comes with the disadvantage that co-occurrence cannot be shown directly in the view anymore but becomes hidden as in node-link diagrams.

5. Reading Fingerprint Matrices

Usually, fingerprint matrices are read line-wise because the fingerprints are easier to compare horizontally than in vertical direction. Reading the matrix requires some training, but there are typical patterns that recur and knowing them eases the analysis significantly. Below we introduce some of these patterns. For space reasons only one row per matrix is displayed for which we exemplify the interpretation.

5.1. Example 1

The figure below exemplifies a character that only plays a role in part of the book (see single person fingerprint). *Jacob* heavily interacts with 3 other characters (2^{nd} to 4^{th} column) of which one (*Alexander*) leaves earlier than the rest. The second time the character is active, he is not part of any group. Furthermore, there is one character in the plot (*Olivia*) that the character *Jacob* never meets. The fluctuations in the fingerprints of *Sophia* and *Emma* show where in the block two co-occurrence lines were connected.

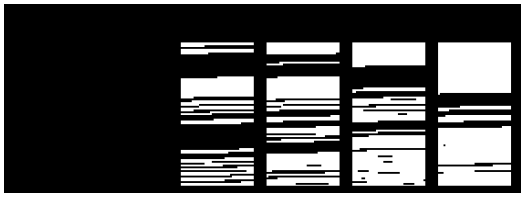


5.2. Example 2

In the second example the character *Martin* is a member of two different groups. It is easy to see that *Martin* switches several times in the course of the novel between the two communities. Furthermore, the different sizes of the colored blocks uncover the different durations of the meetings. The novel concludes with a section in which *Martin* only interacts with *Clara*.

5.3. Example 3

The third example shows a pattern that we observed multiple times in novels, namely that the main protagonist opens the story and the remaining characters are introduced subsequently. The gap in the fingerprints of *Christian* and *Andreas* reveals that there is a passage in the novel in which only *Sabine* and *Svenja* interact with each other (Tobias is not yet present in the story).



6. Case Studies

6.1. Analysis of *Harry Potter and the Philosopher's Stone*

Harry Potter and the Philosopher's Stone is the first novel of a famous fantasy book series by Joanne K. Rowling and was first published in 1997. The whole series is centered around *Harry Potter*, a young boy that discovers during the first volume that he is a wizard and is sent to the "Hogwarts School of Witchcraft and Wizardry" where he soon makes close friends but also some enemies.

Figure 3 shows the fingerprint matrix for the most frequent characters of volume 1. (See supplementary material for a matrix with more characters.) Unsurprisingly, *Harry Potter* is the most frequent character and almost always present during the plot (compare the single person fingerprints at the diagonal). The first row/column shows his co-occurrences with the other characters. Obviously, his close friend *Ron Weasley* almost always stays at his side. Similarly does *Hermione Granger* but she is introduced a bit later and is not as tightly connected to *Harry* at the beginning as *Ron*.

However, before *Harry* meets his new friends *Ron* and *Hermione*, he interacts with three other characters, namely *Albus Dumbledore*, *Rubeus Hagrid*, and *Dudley Dursley*. The fact that most fingerprints in *Dudley*'s row/column are empty reveals that he is only aquainted to few characters. Furthermore, his single person fingerprint shows that he is

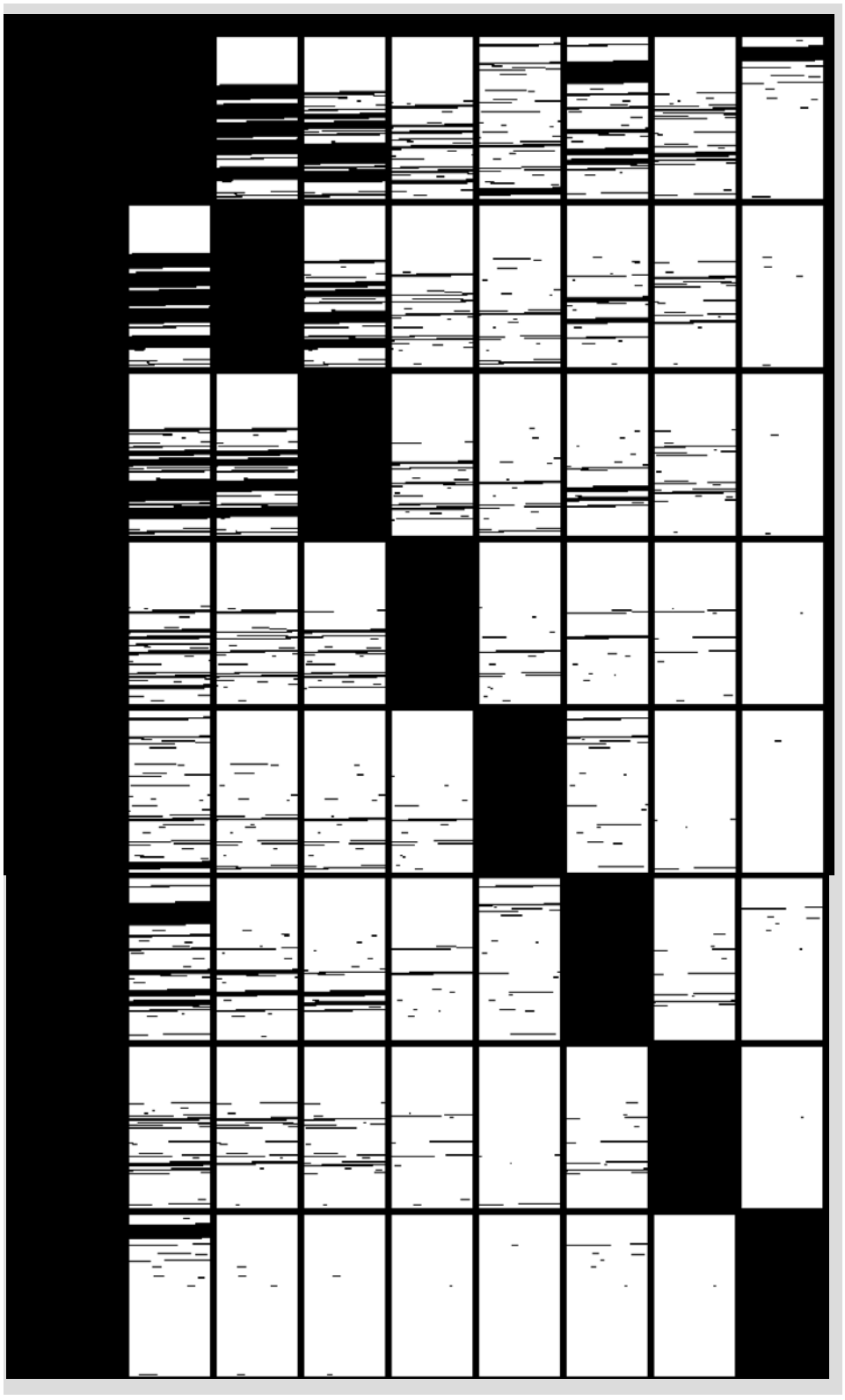


Figure 3: *Fingerprint matrix for the novel Harry Potter and the Philosopher's Stone by Joanne K. Rowling. The matrix was generated with a co-occurrence threshold of 6 and a boosting threshold of 50.*

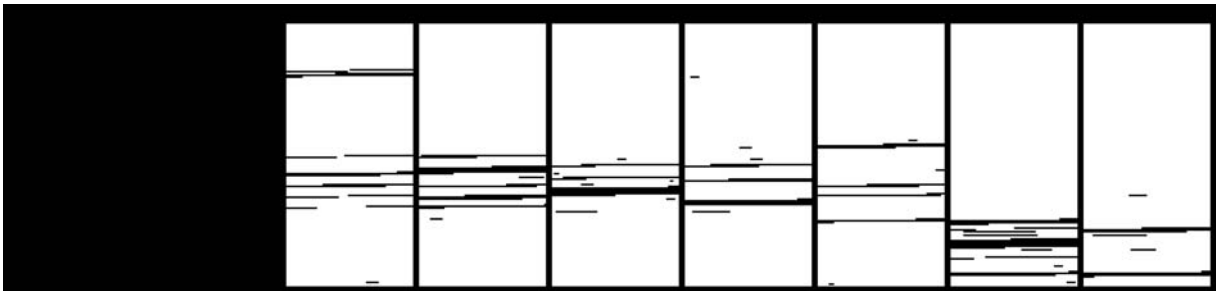


Figure 4: *Excerpt of the fingerprint matrix for the Swedish novel Drottningens juvelsmycke showing the co-occurrences of the main character Tintomara. The figure was generated with a co-occurrence threshold of 6 and a boosting threshold of 50.*

only present at the beginning of the plot. *Dudley Dursley* is *Harry*'s cousin that he lives with since his parents had been killed when he was a baby. The story begins by telling of *Harry*'s life with the *Dursleys* until suddenly *Hagrid*, a half-giant wizard, turns up to inform *Harry* that he has been accepted at Hogwarts School. In this passage we can observe an intense interaction between *Hagrid* and *Harry* in the plot. But also in *Dumbledore*'s fingerprint some colored line snippets can be discerned. It is important to understand that the coloring does not necessarily mean that two characters physically meet in this part of the plot but merely that their names are both mentioned in close proximity. In this case, *Dumbledore* is only mentioned as the author of a letter but does not interact with *Harry* in person.

There are several characters (e.g., *Severus Snape, Draco Malfoy*, and *Rubeus Hagrid*) that meet frequently with *Harry* but do so only for short periods in the novel. To investigate who else is present in these meetings we shift our attention from Harry's row to the rows of the different characters. For instance in *Hagrid*'s row, it becomes obvious that *Ron* and *Hermione* are almost always present when he meets *Harry* in the second part of the book, but *Severus Snape*, one of the tutors of Hogwarts School, misses many encounters.

For comparison in the supplementary material an animated graph can be found (block size = 100 sentences, 10% overlap between neighboring blocks, co-occurrence threshold = 5) as well as two aggregated views.

6.2. Analysis of *The Queen's Tiara*

In our second case study we present an analysis of the Swedish novel *Drottningens juvelsmycke* (English: 'The Queen's Tiara') by Carl Jonas Love Almquist. The novel was published in 1834 and is known as one of Sweden's classics. Its main character *Tintomara* is "a beautiful but sexless androgyne with whom both sexes fall in love" [Goo13]. The plot is set "around the assassination of the 'Theatre *King' Gustav III* on the stage of his own opera house at the masked ball in 1792" [Goo13]. After stealing the Queen's tiara, Tintomara is hidden by a Baroness on her country estate where she meets the daughters of the Baroness, *Amanda* and *Adolfine*. The two were earlier courted by two men, *Ferdinand* and *Clas-Henrik*, who now suspect them to be involved in the king's assassination. All four characters fall in love with *Tintomara* which results in some complications and confusions. At the end of the story, *Tintomara* is killed by one of the men during a performance at the court. [Wik12]

Figure 4 shows the co-occurrences of the protagonist *Tintomara* with the seven most frequent characters of the plot. (See supplementary material for the complete matrix.) From the perspective of the main character *Tintomara* we can recognize three distinct blocks: an introductory block in which Tintomara is almost not present, the second part of the book in which most characters are involved, and finally a closing section that only involves the two characters *Konungen* 'the king' and Gustaf Adolf Reuterholm.

Figure 5 zooms in to an interesting excerpt of the matrix that shows the fingerprints of *Tintomara's* two female lovers *Amanda* and *Adolfine* and the two male lovers *Clas Henrik* and *Ferdinand*. Their fingerprints reveal that all four are already introduced at the beginning of the novel when *Tintomara* is not yet present. Note that since we visualize only co-occurrences and do not look into the semantics of the novel, we do not know at this point if the four get to know each other in this part of the story or if they are mentioned independently of each other.

By taking a closer look at the fingerprints of the four lovers in Figure 4, we can already suppose that the two men do not meet the two women in the second part of the book and vice versa. In the pairwise fingerprints of the four lovers this is now easier to see: in the second half, passages with intensive interactions are only visible between the two female / male characters but not for character pairs with different gender. From the perspective of the plot this likely can be explained with the different identities that *Tintomara* takes which consequently entails the need to keep the two contexts that she appears in separated.

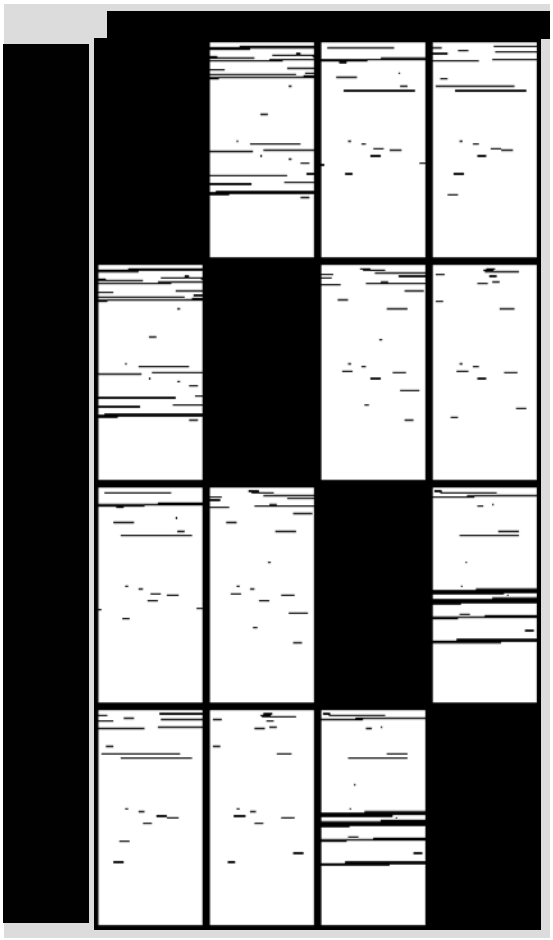


Figure 5: *Excerpt of the fingerprint matrix for the novel Drottningens juvelsmycke 'The Queen's Tiara' showing the co-occurrences of the four lovers of the main character Azouras Lazuli Tintomara. The figure was generated with a co-occurrence threshold of 6 and a boosting threshold of 50.*

6.3. User feedback

To find out how easily accessible our technique is for the intended users, we conducted an informal user study with two literature scholars and four computer scientists of which two are involved in Digital Humanities projects (none of the participants had a background in information visualization). All 6 participants were explained the technique using Figures 1 and 2 in the paper and the examples presented in Section 5. Next, the participants were given the fingerprint matrix shown in Figure 3 and were asked to answer different questions such as which characters are present in the first quarter of the book (Q1), in which order are the characters introduced (Q2), or does character A or B interact longer with character C (Q3). A final question required to order some characters with respect to how often they are a part of the interaction block of two other characters (Q4). Note that Q4 is not easy to answer with the fingerprint matrices because only pairwise interactions are shown and therefore interference is necessary to answer the question. Some participants mentioned explicitly in the questionnaire that this was the most difficult question for them but despite of this 5 of them were able to answer the question correctly which confirms that they understood the technique well. Overall, the visualization was rated as easy to interpret by the participants. What rather surprised us was the fact that most erroneous answers were submitted for Q1. This question is easy to answer if the single person fingerprints are used as a source of information but seemingly the participants focused mainly on the co-occurrence glyphs and forgot about this option.

7. Conclusion

In this paper we presented a technique for visualizing the dynamics of social networks in prose literature. We proposed the usage of fingerprint matrices, which are adjacency matrices that are enhanced with co-occurrence glyphs showing the development of relations across a document. An advantage of fingerprint matrices compared to animated node-link diagrams is that the sequence does not have to be discretized into subsections. Instead, the detected co-occurrences are shown directly in the glyph design, thereby providing the user with the necessary transparency to interpret the data correctly. Furthermore, we apply a boosting technique to increase the visual saliency of the inherent patterns.

Since the relations between characters are determined with a co-occurrence-based measure that does not take the semantics of the text into account, the visual analysis will only be the first step in the literature analysis and is to be followed by closer inspection of interesting passages in the text. We leave it as an issue for future work to develop means to connect the visualization more tightly with the text, potentially also highlighting interesting findings directly in the document. This includes showing the structuring of the text (e.g., the chapter borders) directly in the fingerprints. Furthermore, incorporating additional entities of the plot (e.g., location names) could open the way for answering even more complex analysis questions.

References

[AC07] ABBASI A., CHEN H.: Categorization and analysis of text in computer mediated communication archives using visualization. In *Proc. of the 2007 Conf. on Digital Libraries* (2007), pp. 11–18. 4

[BC02] BRANDES U., CORMAN S. R.: Visual Unrolling of Network Evolution and the Analysis of Dynamic Discourse. In *Proc. of the IEEE Symposium on Information Visualization* (2002). 2

[BE96] BALL T., EICK S. G.: Software Visualization in the Large. *IEEE Computer 29*, 4 (1996), 33–43. 4

[BK10] BORIN L., KOKKINAKIS D.: Literary Onomastics and

Language Technology. In *Literary Education and Digital Learning. Methods and Technologies for Humanities Studies*. IGI Global, 2010, pp. 53–78. 3

[BN11] BRANDES U., NICK B.: Asymmetric Relations in Longitudinal Social Networks. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2283–2290. 2, 4

[BVB*11] BURCH M., VEHLOW C., BECK F., DIEHL S., WEISKOPF D.: Parallel Edge Splatting for Scalable Dynamic Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2344–2353. 2

[BW97] BRANDES U., WAGNER D.: A Bayesian Paradigm for Dynamic Graph Layout. In *Proc. of the 5th Intern. Symposium on Graph Drawing* (1997), pp. 236–247. 2

[CLT*11] CUI W., LIU S., TAN L., SHI C., SONG Y., GAO Z., QU H., TONG X.: Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2412–2421. 2

[CVW09] COLLINS C., VIÉGAS F. B., WATTENBERG M.: Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology* (2009), pp. 91–98. 2

[DZG*07] DON A., ZHELEVA E., GREGORY M., TARKAN S., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proc. of the 16th ACM Conf. on Information and Knowledge Management* (2007), pp. 213–222. 4

[FD00] FEKETE J.-D., DUFOURNAUD N.: Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *Proc. of the 5th ACM Conf. on Digital libraries* (2000), pp. 47–55. 4

[FLE06] FANG S., LWIN M., EBRIGHT P.: Visualization of unstructured text sequences of nursing narratives. In *Proc. of the 2006 ACM Symp. on Applied Computing* (2006), pp. 240–244. 4

[FQ11] FARRUGIA M., QUIGLEY A. J.: Effective Temporal Graph Layout: A Comparative Study of Animation versus Static Display Methods. *Information Visualization 10*, 1 (2011), 47–64. 3

[Goo13] Summary of the novel 'The Queen's Tiara' at Google Books, `http://books.google.de/books/about/The_queen_s_tiara_or_Azouras_Lazuli_Tint.html?id=i3YiAQAAIAAJ`, last accessed on 06/03/13. 8

[Hac09] HACHE B.: *Towards Generic Relation Extraction*. PhD thesis, Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh, 2009. 2

[Hea95] HEARST M. A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proc. of the Conf. on Human Factors in Computing Systems* (1995). 4

[KKEE11] KIM K., KO S., ELMQVIST N., EBERT D. S.: WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora. In *Proc. of the Hawaii Intern. Conf. on System Sciences* (2011), pp. 1–8. 2

[KMOZ08] KEIM D. A., MANSMANN F., OELKE D., ZIEGLER H.: Visual Analytics: Combining Automated Discovery with Interactive Visualizations. In *Proc. of the 11th Intern. Conf. on Discovery Science*, vol. 5254. 2008, pp. 2–14. 4

[KO07] KEIM D. A., OELKE D.: Literature Fingerprinting: A New Method for Visual Literary Analysis. In *IEEE Symposium on Visual Analytics and Technology* (2007), pp. 115–122. 4

[MMBd05] MOODY J., MCFARLAND D., BENDER-DEMOLL S.: Dynamic Network Visualization. *American Journal of Sociology 110*, 4 (2005), 1206–1241. 2

[Mor05] MORETTI F.: *Graphs, maps, trees - abstract models for a literary history*. Verso, 2005. 1

[OJS*11] OELKE D., JANETZKO H., SIMON S., NEUHAUS K., KEIM D. A.: Visual Boosting in Pixel-based Visualizations. *Computer Graphics Forum 30*, 3 (2011). 4

[OKM12] OELKE D., KOKKINAKIS D., MALM M.: Advanced Visual Analytics Methods for Literature Analysis. In *Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2012). 2

[PHG07] PURCHASE H. C., HOGGAN E., GÖRG C.: How Important is the "Mental Map"? - an Empirical Investigation of a Dynamic Graph Layout Algorithm. In *Proc. of the 14th Intern. Conf. on Graph Drawing* (2007), pp. 184–195. 2

[PRY*06] PLAISANT C., ROSE J., YU B., AUVIL L., KIRSCHENBAUM M. G., SMITH M. N., CLEMENT T., LORD G.: Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In *Proc. of the 6th ACM/IEEE-CS Joint Conf. on Digital libraries* (2006), pp. 141–150. 2

[RA00] RIBLER R. L., ABRAMS M.: Using Visualization to Detect Plagiarism in Computer Science Classes. In *Proc. of the IEEE Symposium on Information Vizualization 2000* (2000). 4

[RC11] RYDBERG-COX J.: Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1*, 3 (2011), 1–11. 2

[RES98] ROHRER R. M., EBERT D. S., SIBERT J. L.: The Shape of Shakespeare: Visualizing Text using Implicit Surfaces. In *Proc. of the 1998 IEEE Symposium on Information Visualization* (1998), pp. 121–129. 2

[RTJ*11] REDA K., TANTIPATHANANANDH C., JOHNSON A. E., LEIGH J., BERGER-WOLF T. Y.: Visualizing the Evolution of Community Structures in Dynamic Social Networks. *Computer Graphics Forum 30*, 3 (2011), 1061–1070. 2

[SWS10] STEIN K., WEGENER R., SCHLIEDER C.: Pixel-Oriented Visualization of Change in Social Networks. In *Proc. of the 2010 Intern. Conf. on Advances in Social Networks Analysis and Mining* (2010), pp. 233–240. 2, 4

[TM12] TANAHASHI Y., MA K.-L.: Design Considerations for Optimizing Storyline Visualizations. *IEEE Trans. on Visualization and Computer Graphics 18*, 12 (2012), 2679–2688. 2

[VCPK09] VUILLEMOT R., CLEMENT T., PLAISANT C., KUMAR A.: What's being said near "Martha"? Exploring name entities in literary text collections. In *Proc. of the IEEE Symp. on Visual Analytics Science and Technology* (2009), pp. 107–114. 2

[WFR*01] WOODRUFF A., FAULRING A., ROSENHOLTZ R., MORRSION J., PIROLLI P.: Using thumbnails to search the web. In *Proc. of the SIGCHI Conf. on Human factors in computing systems* (2001), pp. 198–205. 4

[Wik12] Wikipedia page for the novel 'The Queen's Tiara', `http://en.wikipedia.org/wiki/The_Queen's_Tiara`, last accessed on 03/11/12. 8

[WLS*10] WEI F., LIU S., SONG Y., PAN S., ZHOU M. X., QIAN W., SHI L., TAN L., ZHANG Q.: Tiara: a visual exploratory text analytic system. In *Proc. of the Intern. Conf. on Knowledge discovery and data mining* (2010), pp. 153–162. 2

[XS07] XU S., SHIBATA H.: Writing blocks: a visualization to support global revising. In *Proc. of the 19th Australasian Conf. on Computer-Human Interaction: Entertaining User Interfaces* (2007), pp. 61–68. 4

[YEL10] YI J.-S., ELMQVIST N., LEE S.: TimeMatrix: Visualizing Temporal Social Networks Using Interactive Matrix-Based Visualizations. *Intern. Journal of Human-Computer Interaction 26*, 11-12 (2010), 1031–1051. 2, 4