‘Shall I compare thee to a network?’

Visualizing the Topological Structure of Shakespeare’s Plays

Bastian Rieck, *Student Member, IEEE*, and Heike Leitte, *Member, IEEE*

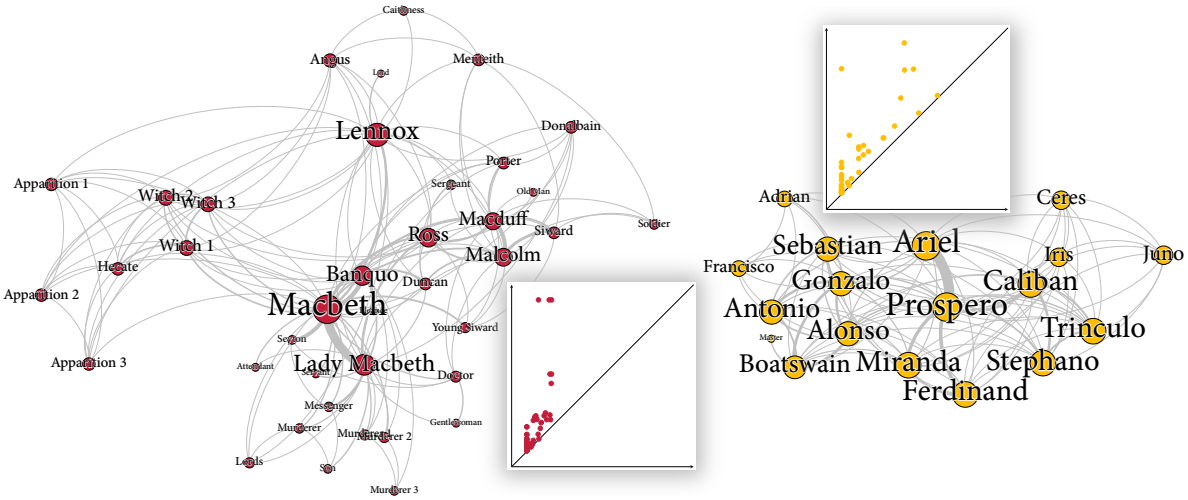


Fig. 1. An illustration of different co-occurrence networks for Shakespeare’s plays. The networks appear to exhibit different structural properties. Assessing their differences is not easy, however. Using persistent homology, we derive a set of structural descriptors (shown as scatterplots in the figure) that permit us to quantify the differences objectively.

Abstract—Many of the plays of William Shakespeare are almost universally known and continue to be played even 400 years after his death. Although the plots of the plays are in general very different, scholars are still discussing similarities in their language, their structure, and many other aspects. In this paper, we demonstrate that visualization approaches may support such an analysis. The presence of machine-readable annotations for each of the plays permits us to construct a set of weighted networks. Every network describes co-occurrence relations between individual characters of a play; its weights may be used to indicate the importance of a connection between two characters, for instance. We subject the networks to a topology-based analysis that permits us to assess their structural similarity. Moreover, we use the dissimilarity values to obtain a topology-based embedding of all the plays. We then proceed to show how features in the dramatic structure of the play manifest themselves in the embedding. This paper is thus a first step towards a more in-depth analysis of the plays, demonstrating the benefits of topology-based visualizations for the digital humanities.

Index Terms—Shakespeare, social network analysis, topology, persistent homology, visualization.

1 Introduction

Even 400 years after his death, Shakespeare remains one of the most prominent and eminent authors. He had a lasting influence on the development of the English language, coining new words such (‘eyeball’, ‘hot-blooded’) or even idioms (‘a foregone conclusion’, ‘heart of gold’). His works thus continue to be of interest to linguists and anglicists. In particular, researchers are interested in finding structural similarities and differences in plays. This can be done by traditional *close reading*—the careful interpretation of single text passages, accounting for all nuances and contexts—or the newly-emerging *distant reading* paradigm that has been pioneered by Moretti [13].

In distant reading, texts are compared among each other using different methods that focus on their relations or structures, for example. This paradigm has aroused a great amount of interest in the visualization community and led to many fruitful collaborations; see e.g. Jänicke [10] for a recent in-depth survey. But what is the appeal of distant reading? In the introduction of his book, Moretti states that “distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.” It is in this spirit that our paper aims to provide a new set of methods for comparing, quantifying, and visualizing structural differences in Shakespeare’s plays. Our method works by extracting *co-occurrence networks*—graphs with edge weights, in which each node represents a particular character in a play—from a tagged corpus of Shakespeare’s plays. We then subject these networks to a novel form of structural analysis by calculating their *persistent homology*. Using persistent homology, we are able to describe a network in terms of its structural features, such as connected components and cycles. Furthermore, we can compare different networks with each other. We demonstrate that this approach can be used effectively to visualize structural differences in plays.

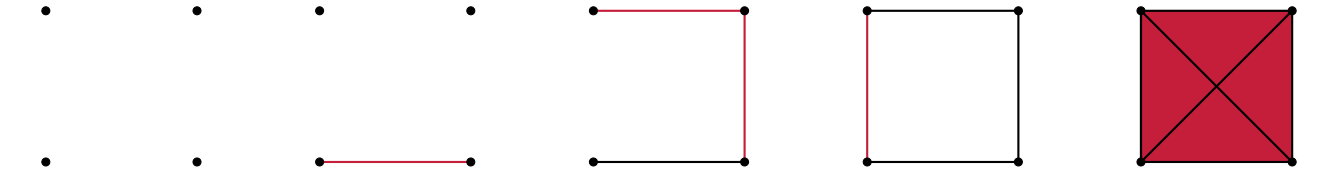
2 Related work

The analysis of social networks is a very established technique in many application disciplines [17]. Typically, these networks are created from real-world networks, such as citation networks or collaboration networks [14]. Our data is somewhat different because we require a separate extraction step.

Many papers concentrate on providing specific visualization strate-

- *Bastian Rieck is with TU Kaiserslautern and Heidelberg University. E-Mail: bastian.rieck@iwr.uni-heidelberg.de.*
- *Heike Leitte is with TU Kaiserslautern. E-Mail: hleitte@cs.uni-kl.de*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx



(a) $\varepsilon = 0.0, \beta_0 = 4, \beta_1 = 0$ (b) $\varepsilon = 0.1, \beta_0 = 3, \beta_1 = 0$ (c) $\varepsilon = 0.5, \beta_0 = 1, \beta_1 = 0$ (d) $\varepsilon = 0.9, \beta_0 = 1, \beta_1 = 1$ (e) $\varepsilon = 1.0, \beta_0 = 1, \beta_1 = 0$

Fig. 2. An illustration of the persistent homology of a graph filtration. We denote the number of connected components by β_0 and the number of cycles by β_1. Since graphs in the filtration need to be nested, we assign every vertex a weight of 0. New edges and triangles are highlighted in red. The scale parameter ε determines for how long structural features persist in the graph. Fig. 4 depicts the corresponding *persistence diagram* $\mathbf{D}$.

gies and graphical user interfaces. The network is either shown directly [8] or in parts [18]. Recent work focuses on visualizing additional attributes in the network that go beyond the co-occurrence relationship [2].

A very common approach for the analysis of networks uses different *centrality measures* [12] to assess the relevance of individual nodes. The distributions of these measures may then be compared to gauge the similarity of networks. We refrain from this approach here because the choice of centrality measure is known to affect the results of the analysis—different measures consider different vertices to be important. Moreover, the accuracy of centrality indices depends on the topology of the graph [4], resulting in biased calculations. We instead focus on describing and summarizing the *structural information* of a network. To this end, we employ *persistent homology*, a method from computational topology [7]. We already demonstrated in previous work that persistent homology is an effective tool for analysing multivariate data sets under multiple aspects [16]. Furthermore, Carstens and Horadam [5] showed that topological features of collaboration networks may be used to distinguish them from e.g. random networks. Their analysis only uses coarse summary statistics, however, whereas our analysis incorporates topological dissimilarity information and topology-based embeddings.

3 Data

For the analysis in this paper, we use a freely-available tagged corpus[1] of Shakespeare's plays. The corpus uses a simple data format that resembles a markup language. Fig. 3 depicts a small example. We parse the play line by line and extract *speakers*, *text*, and relevant *stage directions*, such as a character exiting from stage. Our implementation considers a scene to be the smallest unit of cohesion in the play. Whenever two speakers appear in the same scene, we connect them by an edge. If the edge already exists, we update its weight. We take care to detect speakers that exit from the scene before it is finished, as they contribute less to a scene than a character that remains on stage for the duration of the whole scene.

We use two different weight schemes for the resulting networks: *speech-based* weights and *time-based* weights. Let u and v be two different speakers that co-occur in the same scene. Furthermore, let w_u and w_v be the amount of words used by each speaker and W be the total number of words used in the scene so far. For the *speech-based* weights, we set the weight of edge (u,v) to

$$\mathrm{w}_s(u,v) := \frac{w_u + w_v}{2W}, \tag{1}$$

which ensures that a speaker that only uses a few lines is considered less important than a speaker that has more lines. At the same time, this weight scheme accounts for the fact that speakers that occur in scenes with many lines are generally more important than speakers that occur in scenes with few lines. For example, in the short exchange shown in Fig. 3, the edge between Master and Boatswain has $\mathrm{w}_s(u,v) = 1.0$ because when the Master leaves the scene, he participated in the complete dialog of the scene. The edges connecting the Boatswain to the other characters have larger weights because the Boatswain

```
<ACT 1>
<SCENE 1>
<On a Ship at Sea. A tempestuous noise of thunder &
    ↪ lightning heard.>
<STAGE DIR>
<Enter a Shipmaster and a Boatswain severally.>
</STAGE DIR>
<MASTER> <1%>
        Boatswain!
</MASTER>

<BOATSWAIN> <1%>
        Here, master:  what cheer?
</BOATSWAIN>

<MASTER> <1%>
        Good, speak to the mariners:  fall to't yarely, or
            we run ourselves aground:  bestir, bestir.
<STAGE DIR>
<Exit.>
</STAGE DIR>
</MASTER>
```

Fig. 3. An example of the tagged corpus we use to create networks. We detect speakers, their text, and relevant stage directions.

talks to more characters during the play. This is illustrated in the corresponding network in Fig. 1, right.

We also use a second *time-based* weighting scheme. Here, we use the amount of total time that has elapsed in the play so far. In the tagged corpus, this is encoded as a percentage for every speaker; see Fig. 3. Following the notation introduced above, we set the weight of edge (u,v) to

$$\mathrm{w}_t(u,v) := \max(t_u, t_v), \tag{2}$$

where t_u and t_v denote the time at which characters u and v appeared for the first time in the play. For example, the edge between Master and Boatswain has a weight of 1, because both characters co-occur when about 1% of the play has elapsed. This weight scheme is more coarse but it permits us to compare the temporal evolution of different plays with each other, i.e. whether characters co-occur at similar times.

Visualization & post-processing. We use the open-source tool *Gephi* [1] and the ForceAtlas2 [9] algorithm for visualizing the graphs. Fig. 1 shows some example graphs in which the thickness of an edge denotes its weight. These visualizations show that our weight schemes are capable of displaying structural differences between different networks. We use them also for manually correcting the resulting networks because our parser implementation fails to create edges between minor characters in some of the plays. To encourage further research and critique, we make the original and the corrected networks as well as the parser code freely available[2].

[1] http://lexically.net/wordsmith/support/shakespeare.html

[2] https://github.com/Submanifold/Shakespeare

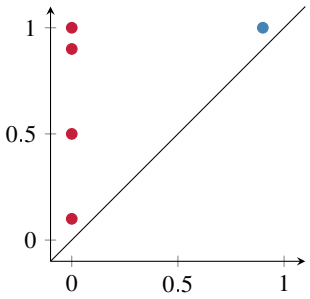


Fig. 4. A persistence diagram $\mathbf{D}$. The structural features, i.e. connected components ● and cycles ●, obtained during the graph filtration depicted in Fig. 2 are summarized here. Persistence diagrams afford robust distance metrics that quantify topological similarity between two graphs.

4 Persistent Homology

Persistent homology is a method from *computational topology* [7] that is often used to detect topological features in high-dimensional data sets. In the case of social networks, which are a special class of graphs, these features are well-known and comprise *connected components* (dimension 0) and *cycles* (dimension 1). Connected components indicate clusters of characters in a graph, while cycles indicate small cliques, i.e. characters that often appear together with a similar importance. These features are known to represent salient properties of a network [5]. The basic idea of persistent homology is to assess a graph $\mathbf{G}$ at multiple scales in order to increase the amount of structural information that is used. To perform this multi-scale assessment, a scalar function on the graph—such as a *weight function*—is required. In the following we refer to a $\mathbf{G}(\varepsilon)$ as the subgraph of $\mathbf{G}$ in which an edge (u,v) is only kept if $\mathrm{w}(u,v) \leq \varepsilon$. We have a natural nesting relation between graphs for increasing weights, i.e. $\mathbf{G}(\varepsilon) \subseteq \mathbf{G}(\varepsilon')$ for $\varepsilon \leq \varepsilon'$. As a consequence, given a series of thresholds $\varepsilon_0, \ldots, \varepsilon_k$, we obtain a sequence of nested graphs,

$$\emptyset \subseteq \mathbf{G}(\varepsilon_0) \subseteq \mathbf{G}(\varepsilon_1) \subseteq \cdots \subseteq \mathbf{G}(\varepsilon_{k-1}) \subseteq \mathbf{G}(\varepsilon_k) \subseteq \mathbf{G}, \tag{3}$$

which we refer to as a *graph filtration*. Notice that the two weight schemes we defined above are valid graph filtrations in this sense. We also add all triangles, i.e. 3-cliques, to the graph. Adding these triangles has the effect of removing cycles from the graph. This permits us to measure at which values of the weight parameter ε a cycle starts to disappear again because it has been 'filled in'. Fig. 2 depicts an example of how to calculate persistent homology of a graph filtration. At every value of the weight parameter ε, we analyse a different 'snapshot' in the graph and count the number of connected components β_0 as well as the number of cycles β_1. The number of structural features may change between snapshots, as features may be 'created' or 'destroyed'. For example, the inclusion of an edge may destroy a connected component because it merges with another connected component. We summarize these changes of structural features in an auxiliary visualization, the *persistence diagram* $\mathbf{D}$. If a feature has been created in a snapshot for $\varepsilon = \varepsilon_i$ and destroyed at $\varepsilon = \varepsilon_j$, we add the point $(\varepsilon_i, \varepsilon_j)$ to the persistence diagram $\mathbf{D}$. In the graph filtration shown in Fig. 2, for example, all connected components are created at $\varepsilon = 0$. At $\varepsilon = 0.1$, the inclusion of an edge merges two components, so we add the point $(0, 0.1)$ to $\mathbf{D}$. Fig. 4 shows the full persistence diagram for this example.

We refer to the quantity $\mathrm{pers}(\varepsilon_i, \varepsilon_j) := |\varepsilon_j - \varepsilon_i|$ as the *persistence* of the corresponding feature. The persistence of a point is an indicator of its relevance. Features that persist over a long range of the weight parameter ε are considered to be important, while features that persist only over a short range are often considered to be noise. In the persistence diagram $\mathbf{D}$, $\mathrm{pers}(\cdot)$ is represented by the distance to the diagonal. The closer a point is to the diagonal, the less important its corresponding feature. Following Cohen-Steiner et al. [6], we add the point $(0, \max \varepsilon)$ to our persistence diagrams to ensure that all connected components are represented in the $\mathbf{D}$.

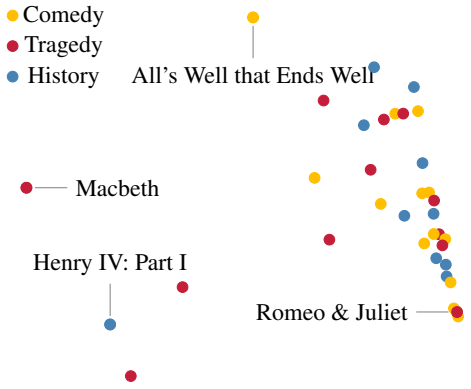


Fig. 5. An embedding of all plays according to the speech-based filtration. Some plays are structurally extremely different from the remaining plays, which results in them being treated as outliers. For layout reasons, we do not show all labels.

Comparing persistence diagrams. The appeal of persistent homology is that there are robust distance metrics between persistence diagrams that permit assessing the topological similarity of graphs. In this paper, we use the *Wasserstein distance* W_2. Given two persistence diagrams $\mathbf{X}$ and $\mathbf{Y}$, it is defined as

$$\mathrm{W}_2(\mathbf{X}, \mathbf{Y}) := \left(\inf_{\eta \colon \mathbf{X} \to \mathbf{Y}} \sum_{x \in \mathbf{X}} \|x - \eta(x)\|_\infty^2 \right)^{\frac{1}{2}}, \tag{4}$$

where $\eta \colon \mathbf{X} \to \mathbf{Y}$ denotes a bijection and $\|\cdot\|_\infty$ the *maximum norm*. The Wasserstein distance thus measures the amount of transformations required to transform one persistence diagram into another one. Since the cardinality of both diagrams is different in most cases, we assume that $\mathbf{X}$ and $\mathbf{Y}$ also contain the orthogonal projections of their points to the diagonal. The bijection η may then send a point in one diagram to its projection onto the diagonal, thereby indicating an unmatched topological feature.

Implementation & performance Persistent homology for graphs can be efficiently calculated, requiring only a single pass through the graph. This has a complexity of $\mathcal{O}(n \log n)$, where n refers to the number of edges in the graph. The Wasserstein distance W_2 requires calculating maximum weighted matchings in bipartite graphs [7, pp. 229–236]. The complexity of this calculation is $\mathcal{O}(m^3)$, where m refers to the number of points in the persistence diagram. This is not yet prohibitive for our persistence diagrams, but novel algorithms [11] make the Wasserstein distance calculations applicable even for very large persistence diagrams.

5 Results

We now briefly discuss the results for the two filtrations we defined earlier.

Speech-based filtration

First, we calculate persistent homology for the *speech-based* filtration to obtain the pairwise Wasserstein distances between the resulting persistence diagrams. Using *metric multidimensional scaling* [3], we obtain an embedding of all plays in $\mathbb{R}^2$ (Fig. 5). In this embedding, spatial proximity indicates topological—i.e. *structural*—similarity. We follow the classification into *comedy* ●, *tragedy* ●, and *history* ● as mentioned in the *First Folio* of Shakespeare's plays. In the embedding, several plays are conspicuous and get our attention. The tragedy Romeo & Juliet, for instance, is surrounded by several comedies and histories. This means that its *structure* resembles that of a comedy more than it that of a tragedy. Some essays in literary criticism, e.g. Snyder [19], already stated similar results. It is interesting to see that content-based analysis can be supplemented by our quantifiable structural analysis of the 'constellation of characters' and the amount of their

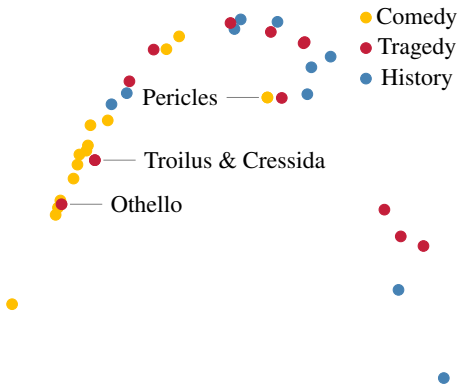


Fig. 6. An embedding of all plays according to the time-based filtration. There is a clearly-visible strand of 'comedic' plays, with some interesting outliers. For layout reasons, we do not show all labels.

interaction in the play. Moreover, we observe that MACBETH, another tragedy, is placed well apart from all other plays. This is because it features more connected components, i.e. more 'character clusters', with higher persistence values, as well as a larger amount of cycles of high persistence—indicating that there are more subplots in MACBETH than in other plays. Fig. 1 depicts two compressed persistence diagrams. We can see that in contrast to the diagram for A MIDSUMMER NIGHT'S DREAM, the diagram for MACBETH contains data points that are removed from the diagonal.

Similar observations apply to ALL'S WELL THAT ENDS WELL, a comedy that is structurally different from the other plays—a fact that has already been studied using *close reading* [21]. Furthermore, HENRY IV: PART I is set apart from all other plays, in particular from the remaining histories of the 'Henriad', due to its highly-complex structure comprising numerous intertwined subplots and many smaller figure constellations that result in a large amount of topological activity. We did not find a comprehensive justification for this phenomenon.

Time-based filtration

Fig. 6 shows the embedding of all plays according to the time-based filtration. We observe that most comedies are 'lumped together' in a longer elongated structure. This implies that there is a relatively typical order in which different characters are introduced in a comedic play. The most notable outlier is the comedy PERICLES, PRINCE OF TYRE, which appears to be rather atypical for the remaining comedies. Interestingly, this is a play to which Shakespeare only *contributed* as a co-author.

Furthermore, we can see that two tragedies, OTHELLO and TROILUS & CRESSIDA are situated near the cluster of comedies. The internal chronology of these plays is thus most similar to that of a typical comedy. Interestingly, this fact has already been noticed by Shakespeare scholars, leading to the definition of a Shakespearean *problem play*. A problem play is play with an ambiguous tone that cannot be easily categorized. Tillyard [21], for instance, considers TROILUS & CRESSIDA to be the epitome of a problem play. Likewise, Teague [20] writes about how OTHELLO features a 'fundamentally comic structure'.

The time-based structural analysis of a play can give further weight to these traditional text-based studies. It also serves to highlight the difficulty in classifying a play by its internal chronological structure alone—there are no well-separated clusters between the different genres.

6 Conclusion

In this paper, we presented a novel method for quantifying structural information in the plays of William Shakespeare. Our method uses topological concepts to analyse weighted co-occurrence networks of a play. Using two different strategies for calculating edge weights in the networks, we were able to assess dissimilarities between plays from two different viewpoints. We demonstrated that our method is capable of visualizing interesting patterns with respect to the structure of a play. Our analysis remained relatively superficial because our goal was to demonstrate that topology-based techniques are capable of detecting salient structures in these networks.

We hope that our paper stimulates further research in that regard, as there are multiple aspects that may increase the expressive power. For example, it would be interesting to obtain weights that are based on the emotional content or setting of scene. This would tie in with recent work by Reagan et al. [15] on emotional arcs in stories. Moreover, the visualizations we showed here could be extended to depict the chronology of Shakespeare's own work. This could result in a visualization that shows changes in structural style over time.

Acknowledgments

The authors wish to thank Ingo Kleiber for inspiration and for providing the initial version of the co-occurrence extraction code.

References

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Web and Social Media*, 2009.
[2] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J.-D. Fekete. GraphDice: A system for exploring multivariate social networks. *Computer Graphics Forum*, 29(3):863–872, 2010.
[3] I. Borg and P. J. F. Groenen. *Modern multidimensional scaling*. Springer Series in Statistics. Springer, New York, NY, USA, 2nd ed., 2005.
[4] S. P. Borgatti and M. G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 2006.
[5] C. J. Carstens and K. J. Horadam. Persistent homology of collaboration networks. *Mathematical Problems in Engineering*, 2013, 2013.
[6] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Extending persistence using Poincaré and Lefschetz duality. *Foundations of Computational Mathematics*, 9(1):79–103, 2009.
[7] H. Edelsbrunner and J. Harer. *Computational topology: An introduction*. American Mathematical Society, Providence, RI, USA, 2010.
[8] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization*, pp. 32–39, 2005.
[9] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. FORCEATLAS2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, 9(6):1–12, 2014.
[10] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. In R. Borgo, F. Ganovelli, and I. Viola, eds., *Eurographics Conference on Visualization (EuroVis) — STARs*. The Eurographics Association, 2015.
[11] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. In M. Goodrich and M. Mitzenmacher, eds., *Proceedings of the 18th Workshop on Algorithm Engineering and Experiments* (ALENEX), pp. 103–112, 2016.
[12] T. Martin, X. Zhang, and M. E. J. Newman. Localization and centrality in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 90:052808, 2014.
[13] F. Moretti. *Graphs, maps, trees: Abstract models for a literary history*. Verso, London, England, 2007.
[14] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.
[15] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds. The emotional arcs of stories are dominated by six basic shapes. *CoRR*, abs/1606.07772, 2016.
[16] B. Rieck and H. Leitte. Exploring and comparing clusterings of multivariate data sets using persistent homology. *Computer Graphics Forum*, 35(3):81–90, 2016.
[17] J. Scott. *Social Network Analysis*. SAGE Publications, Inc., 2013.
[18] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):733–740, 2006.
[19] S. Snyder. Romeo and juliet: Comedy into tragedy. *Essays in Criticism*, 20(4):391–402, 1970.
[20] F. Teague. *Othello* and new comedy. *Comparative Drama*, 20(1):54–64, 1986.
[21] E. M. W. Tillyard. *Shakespeare's problem plays*. Chatto and Windus, London, 1949.