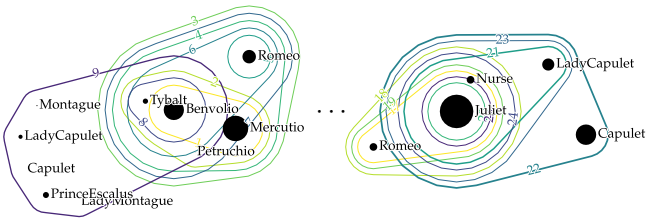
All the World's a (Hyper)Graph: A Data Drama

Corinna Coupette[1], Jilles Vreeken[2], and Bastian Rieck[3,4]
[1]Max Planck Institute for Informatics [2]CISPA Helmholtz Center for Information Security
[3]Institute of AI for Health, Helmholtz Munich [4]Technical University Munich

We introduce HYPERBARD, a dataset of diverse relational data representations derived from Shakespeare's plays [1]. Our representations range from simple graphs capturing character co-occurrence in single scenes to hypergraphs encoding complex communication settings and character contributions as hyperedges with edge-specific node weights. By making multiple intuitive representations readily available for experimentation, we facilitate rigorous representation robustness checks in graph learning, graph mining, and network analysis, highlighting the advantages and drawbacks of specific representations. Leveraging the data released in HYPERBARD, we demonstrate that many solutions to popular graph mining problems are highly dependent on the representation choice, thus calling current graph curation practices into question. As an homage to our data source, and asserting that science can also be art, we present all our points in the form of a play.



(a) Act III, Scene I (b) Act III, Scene V

Fig. 1: Interactions between named characters in *Romeo and Juliet*, modeled as a hypergraph resolved at the stage-group level. Edge labels denote stage groups, edge colors indicate edge order, and node sizes and edge widths are proportional to the number of spoken lines. Unlike classic co-occurrence graphs, fine-grained hypergraphs retain, e.g., the crucial fact that Juliet's parents never meet Romeo in Act III, Scene V.

The Story. *Induction, Scene I.* Confronted by REVIEWER, AUTHORS explain their first contribution. *Act I, Scene I.* CREATURE gets drawn into the Community by SENIOR RESEARCHER and TUTOR. Welcomed by PROFESSOR, they sign their PhD contract. *Act I, Scene II.* CREATURE quarrels with their new role. They meet COLLEAGUE, their office mate, and three DEADLINES, introduced by PROFESSOR. They submit to FIRST DEADLINE. *Act I, Scene III.* CREATURE dreams of HYPERBARD, a faun caring for raw data, and GRAPH, one of their spirits. They discuss how to obtain insights from raw data via transformations, and that each raw data point permits several relational representations. *Act II, Scene I.* CREATURE converses with COLLEAGUE, PROFESSOR, and SENIOR RESEARCHER over lunch. They ask COLLEAGUE about the provenance of graph data used in the Community, and they learn about graph data repositories. *Act II, Scene II.* CREATURE revisits their dream. They identify semantic mapping, granularity, and expressivity as the dimensions in which several graph representations of the same raw data may differ. *Act II, Scene III.* CREATURE secretly observes COLLEAGUE as they mechanically prepare a graph dataset and produce a datasheet in the process. *Act II, Scene IV.* Confused and depressed by the practices they witness in the Community, CREATURE attempts suicide. *Act II, Scene V.* Outside the Community, CREATURE is cared for by GRAPH and HYPERBARD. Together, the three of them develop the graph and hypergraph representations of Shakespeare's plays included in the HYPERBARD dataset. *Act III, Scene I.* CREATURE gets haunted by the three DEADLINES, who remind them of their ignoble academic incentives. They contemplate quitting their PhD. *Act IV, Scene I.* Accompanied by GRAPH and HYPERBARD, CREATURE returns to the Community. They meet PROFESSOR, who calls CREATURE into their office and demands that HYPERBARD leaves. *Act IV, Scene II.* From PROFESSOR, CREATURE learns that their paper got accepted. *Act IV, Scene III.* In the absence of CREATURE, HYPERBARD and GRAPH try to convey their message that representations matter to COLLEAGUE. PROFESSOR and CREATURE return, and PROFESSOR orders COLLEAGUE to eliminate HYPERBARD. *Act V, Scene I.* Having cremated HYPERBARD, COLLEAGUE pours their ashes onto the graph dataset prepared earlier. GRAPH mourns the death of their sovereign and sketches its implications. *Act V, Scene II.* CREATURE wrestles with their experience in the Community. Instead of leaving in silence, they decide to tell their own story.

The Dataset. The HYPERBARD dataset comprises 666 graphs and hypergraphs: 18 relational representations for each of 37 plays by William Shakespeare. From the TEI Simple XMLs provided by Folger Digital Texts, for each play, we derive 6 hypergraphs, 6 clique expansions (i.e., interaction graphs), and 6 star expansions (i.e., bipartite graphs) that differ along 3 dimensions: *semantic mapping*, *granularity*, and *expressivity*. As we show for *Romeo and Juliet*, these representations emphasize different aspects of the underlying raw data, and they yield widely varying results even for simple measurements of character importance. Thus, HYPERBARD *enables* and *demonstrates the need for* research on how representation choices impact the outputs and performance of graph learning, graph mining, and network analysis methods.

The Critique. *The Community* is designed as a microcosm of *our research community*, including all levels of academic seniority as well as common supporting roles. The characters *inside* the Community exhibit cognitive, behavioral, and interaction patterns that frequently afflict people with corresponding roles in our community. The characters *outside* the Community appear as their antidotes, challenging the status quo and engaging in free-spirited scientific inquiry. As the play progresses, CREATURE gets caught up between both worlds, and we witness the force of community dynamics acting upon individuals that do not fit in.

[1] C. Coupette, J. Vreeken, and B. Rieck, *All the World's a (Hyper)Graph: A Data Drama*, Submitted (2022/23). Preprint available at https://arxiv.org/abs/2206.08225.