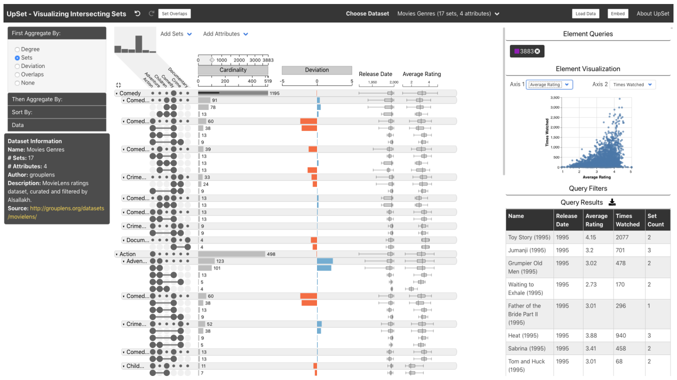
UpSet 2: From Prototype to Tool

Kiran Gadhave*
University of Utah

Hendrik Strobelt†
IBM Research AI

Nils Gehlenborg‡
Harvard Medical School

Alexander Lex§
University of Utah



UpSet - Visualizing Intersecting Sets

Choose Dataset: Movies Genres (17 sets, 4 attributes)

First Aggregate By:

○ Degree
● Sets
○ Deviation
○ Overlaps
○ None

Then Aggregate By:

Sort By:

Data

Dataset Information
Name: Movies Genres
# Sets: 17
# Attributes: 4
Author: grouplens
Description: MovieLens ratings dataset, curated and filtered by Alsallakh.
Source: http://grouplens.org/datasets/movielens/

Element Queries

3883

Element Visualization

Axis 1: Average Rating — Axis 2: Times Watched

Query Filters

Query Results

Name	Release Date	Average Rating	Times Watched	Set Count
Toy Story (1995)	1995	4.15	2077	2
Jumanji (1995)	1995	3.2	701	3
Grumpier Old Men (1995)	1995	3.02	478	2
Waiting to Exhale (1995)	1995	2.73	170	2
Father of the Bride Part II (1995)	1995	3.01	296	1
Heat (1995)	1995	3.88	940	3
Sabrina (1995)	1995	3.41	458	2
Tom and Huck (1995)	1995	3.01	68	2

Figure 1: The refactored UpSet 2. Key improvements address the ease of adoption, sharing, and integration with other tools.

Abstract

UpSet is a technique for visualization and analysis of sets and their intersections which was introduced at InfoVis 2014. The technique visualizes the elements and their set memberships in a matrix layout along with an aligned bar chart to display intersection sizes. As the approach provides a more accurate representation of the size of set intersections and scales better with respect to the number of sets than other approaches, UpSet plots are frequently used in papers in the biomedical domain. However, we believe that this popularity is mostly due to an R version that generates UpSet figures, and not because of our interactive JavaScript implementation that was published with the original paper. Why is the R version more popular? We believe it has multiple reasons, but one of them is that the original tool is an academic prototype, with the usual shortcomings for practical use. This poster presents a new version of UpSet—a tool that was developed to address requests for additional functionality over the original implementation based on user feedback. We describe in detail which changes we undertook to improve the paper prototype into a tool. The new version adds features to improve adoption, improve sharing insights, and allow customization. For example, it allows embedding of the tool in webpages and supports tracking of provenance to provide a undo and redo functionality. The new implementation also can be used as a JavaScript/TypeScript library, which makes it easy to integrate UpSet in larger systems.

Index Terms: Human-centered computing—Visualization—Visualization techniques

1 Introduction

Visualizing sets and their intersections is an essential task in the analysis of set data. A popular method to visualize set intersections are Venn diagrams. However, Venn diagrams work well only for visualizing intersection of three to four sets. UpSet [4] addresses this issue by visualizing sets and their intersections in a matrix layout, where each row corresponds to an intersection, and each column to a set (see Figure 1). The size of intersections and other attributes are encoded by bar charts and box plots next to the matrix. The tool also supports aggregation of set intersections by degree, sets, deviation, or overlaps. These aggregations enable analysts to investigate complex relationships. The basic UpSet view is supported by views showing properties of elements in a table or in supplementary visualizations.

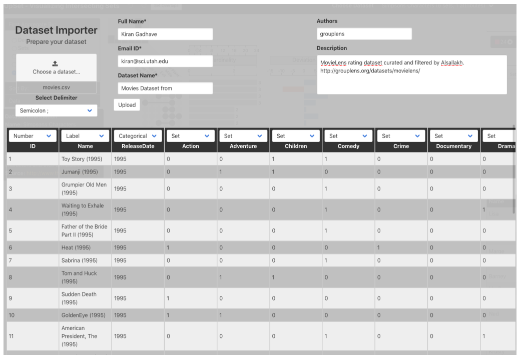
Conway et al. have released an R implementation of the tool called UpSetR [1], which generates static UpSet plots. The design is modified slightly to account for the difference of static, paper and print focused use cases, and various advanced features are missing from the R version. Using UpSetR requires scripting in R and passing all the configuration as parameters to the function provided by the library or using a Shiny app. Other groups not affiliated with the original authors have created multiple UpSet versions for Python.

UpSet plots have become quite popular in the biomedical domain. The two papers have accumulated more than 450 citations in the five respectively two years since publication. UpSetR has been downloaded about 150k times, however, the interactive JavaScript version has been accessed by only about 16k users since November 2015. We speculate the main reason UpSetR is more popular is that the R version fits nicely into established bioinformatics workflows and produces publication-ready figures.

However, we also believe that another factor plays a role: the original UpSet implementation works reasonably well, but still is an academic prototype, with the usual shortcomings: limited input and output formats, no undo/redo, no ability to create publication-quality figures, and no ability to integrate with other workflows.

To address these shortcomings, we have developed a new, web-based version of UpSet. UpSet 2, available at `https://vdl.sci.utah.edu/upset2/` is a re-write from scratch using the latest web technologies, largely following the original design, but adding data upload, undo/redo, and data sharing. To enable the integration of UpSet into larger projects, we provide UpSet 2 as a library exposing an easy to use API. Finally, to allow users to disseminate the interactive UpSet 2 plots, we provide the ability to embed the plots in arbitrary websites, where they can serve as interactive figures. We believe that such ‘interactive figures’ that also capture analysis provenance [3] are an important development for academic publishing, as is evident by the emergence of platforms such as `distill.pub` and e-life’s reproducible articles [2].

*e-mail: kiran@sci.utah.edu
†e-mail: hendrik@strobelt.com
‡e-mail: nils@hms.harvard.edu
§e-mail: alex@sci.utah.edu



Dataset Importer

Prepare your dataset

Choose a dataset...

movies.csv

Select Delimiter: Semicolon ;

Full Name*: Kiran Gadhave

Email ID*: kiran@sci.utah.edu

Dataset Name*: Movies Dataset from

Upload

Authors: grouplens

Description: MovieLens rating dataset curated and filtered by Alsallakh. http://grouplens.org/datasets/movielens/

Number ID	Label Name	Categorical ReleaseDate	Set Action	Set Adventure	Set Children	Set Comedy	Set Crime	Set Documentary	Set Drama
1	Toy Story (1995)	1995	0	0	1	1	0	0	0
2	Jumanji (1995)	1995	0	1	1	0	0	0	0
3	Grumpier Old Men (1995)	1995	0	0	0	1	0	0	0
4	Waiting to Exhale (1995)	1995	0	0	0	1	0	0	1
5	Father of the Bride Part II (1995)	1995	0	0	0	1	0	0	0
6	Heat (1995)	1995	1	0	0	0	1	0	0
7	Sabrina (1995)	1995	0	0	0	1	0	0	0
8	Tom and Huck (1995)	1995	0	1	1	0	0	0	0
9	Sudden Death (1995)	1995	1	0	0	0	0	0	0
10	GoldenEye (1995)	1995	1	1	0	0	0	0	0
11	American President, The (1995)	1995	0	0	0	1	0	0	1

Figure 2: Data upload interface with support for splitting data, assigning data types to columns, and selecting columns which contain set information.

2 From Prototype to Tool

UpSet 2 improves on UpSet by improving in three areas: easing adoption, easing sharing, and enable integration. We discuss specific improvements along these three dimensions.

2.1 Ease Adoption: Data Upload

If a tool makes it difficult to import a user's data, it is likely that they will look in other places for similar functionality. The original UpSet required users to host datasets on a public server, and write a JSON file specifying the contents of the data. To make the tool easy to adopt, we added a data upload functionality to UpSet 2 that makes evaluation of its fitness by new users and adoption easier. Fig. 2 shows the new data upload module. It allows interaction with a visual, tabular representation of a dataset. The UI enables the users to split the data, to assign data types to columns, and to select columns which encode information about set membership. The data is uploaded to a public server and can be made available publicly to other users.

2.2 Ease Adoption: Provenance Integration

Enabling undo/redo has been a mainstay recommendation for visualization tool for decades. Yet very few academic tools support it. Provenance tracking is now built into the web-based tool and tracks all user actions like aggregations, sorting, addition/removal of sets, selections, and queries. Provenance tracking allows for undo and redo of user operations. It is also possible to download the provenance data and share one's analysis session. This enables reproducibility of analysis steps, which is an important feature and eases sharing of research, validation of analysis, and recording thought processes.

2.3 Ease Sharing: Embedding UpSet 2

As web-based articles and blog posts are becoming more popular for scientific dissemination, there is an opportunity for using interactive figures. UpSet 2 enables this by providing an easy way to include an interactive plot on a website using an embed code, in the style of youtube or twitter embedding snippets.

Using the embed feature requires minimal programming/scripting skills and minimal configuration. Most blogging frameworks provide features to embed iframes, hence UpSet 2 provides the embed feature in an iframe snippet. The embed pop-up provides a code snippet which the users can paste in existing HTML documents.

As the full UpSet is design for full-screen use, UpSet 2 can be customized to only show selected views, which is more suitable for inclusion in a website. Similarly, filters, and other on-screen elements can be removed on demand. A demonstration of an embedded UpSet 2 plot can be found at `https://caleydo.org/tools/upset/#upset2`.

2.4 Ease Sharing: Download Selection

One of the most commonly requested features for the prototype was a download function for interactions and aggregates. Analysts want to use UpSet to identify a particular subset and then continue processing this subset in another tool. UpSet 2 introduces a download feature. Selection of intersections, aggregates, or queries in the matrix view list the corresponding elements in the element table. The content of this table can be downloaded as a text file maintaining the same delimiters as the original file. Combined with aggregates and queries, this allows for downloading subsets of data matching complicated set relations and can facilitate the use of UpSet in a larger analysis pipeline.

2.5 Enable Integration: Upset as Library

For users with advanced requirements, including visualization developers and researchers, Upset 2 is also available as a Javascript/Typescript library and exposes a simple API with one function. The function allows customization of the visualizations, filters, dataset, and size. The library version allows for deeper integration of the technique with a visualization tool or dashboard. Developers can use libraries like D3 to customize the look of UpSet plots.

3 Implementation

UpSet 2 is written entirely in Typescript and provides type support for the library component. Using Typescript allows for compile-time checks and makes it easy to develop the project further. The data is stored on a server written in NodeJS and hosted on an Amazon EC2 instance. It is a public server, and anyone using the tool can access all the uploaded datasets. Files with the same combination of dataset name and email id are not allowed. UpSet 2 is open source and available at `https://github.com/visdesignlab/upset2`.

4 Conclusion and Discussion

We re-wrote UpSet to make the leap from an academic prototype to a production-ready tool. We hope that our improvements will make it easier for analysts to get their data into upset, to export the relevant subsets, to share what they found in interactive figures, and to integrate (parts of) UpSet in larger systems. While we don't expect UpSet 2 to be quite as widely adopted as UpSetR, we hope that our improvements will lead to a wider adoption.

However, we plan to furter improve UpSet 2. The current version of the tool expects the data to be in a binary matrix. We plan to integrate other data formats, such as explicit lists of item memberships in sets. We also plan on generating code for UpSetR based on the configuration of an UpSet plot, resulting in reproducible, publication quality figures for users with an R workflow.

References

[1] J. R. Conway, A. Lex, and N. Gehlenborg. UpSetR: An R Package for the Visualization of Intersecting Sets and their Properties. *Bioinformatics*, 33(18):2938–2940, 2017. doi: 10.1093/bioinformatics/btx364

[2] Giuliano Maciocci, Michael Aufreiter, and Nokome Bentley. Introducing eLife's first computationally reproducible article. https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article, Feb. 2019.

[3] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit. From Visual Exploration to Storytelling and Back Again. *Computer Graphics Forum*, 35(3):491–500, 2016. doi: 10.1111/cgf.12925

[4] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec. 2014. doi: 10.1109/TVCG.2014.2346248