

Chord2DS: An Extension to Chord Diagram to Show Data Elements from Two Heterogeneous Data Sources

Shah Rukh Humayoun and Likhitha Brahmadevara

Department of Computer Science, San Francisco State University, USA
humayoun@sfsu.edu, lbrahmadevara@mail.sfsu.edu

Abstract

The standard Chord diagram, a radial layout, shows data elements in a circular fashion from one data source. In this paper, we propose an extension to the standard Chord diagram to show data elements from two heterogeneous data sources into one single diagram. The main Chord diagram is used for showing data elements and the relations between them from one data source, while we use an outer layer to show data elements from the second data source. The relationships between data elements from both data sources are shown through visual cues. The proposed solution uses space efficiently compared to using multiple diagrams in the scenarios of two heterogeneous data sources.

CCS Concepts

• **Human-centered computing** → **Information visualization; Graphical user interfaces;**

1. Introduction

Researchers have suggested that radial (circular) layouts produce compact visualizations and use space efficiently, as they support a larger data domain on a squared area compared to rectangular or square layouts [KSB*09] [Krz]. Eye movement is also encouraged to proceed along the curved lines of circular layouts, which results in better exploration of underlying data [Krz]. Furthermore, they are useful to show the flow on relations between data element pairs more intuitively [KSB*09].

The Chord diagram is a radial layout which was initially popularized by The New York Times to show the relations between Genomes using the Circos package [KSB*09] [TNYT07]. Later on Chord diagram has been used and extended by other researchers. Few examples are: In [QAML17], Wi et al. used Chord diagram to plot China's inter-provincial migration during the period of 2010–2015. In [ZSB*17], Zoun et al. used Chord diagram to explore microorganisms that live in complex microbial communities. In [Jal16], Jalali used various Chord diagrams for process mining in social network analysis. In [HAAE17a], Humayoun et al. extended Chord diagram to visualize an ego network and the associated social circles. In [HAAE17b], Humayoun et al. provided an extended Chord diagram to show the important keywords regarding tweet data of an event and the relations between these keywords. While in [HBA18], Humayoun et al. used Chord diagram to show the relations between bi-categorical dimensional data where each categorical dimension uses half of the circle.

In standard Chord diagrams, data elements (also called *nodes* or *arcs*) are arranged in a circular fashion and the relations (also

called *chords*) are drawn between these nodes. In general, a standard Chord diagram is used to represent data elements and the relations between them from one data source (dataset). However, sometimes we may have two heterogeneous data sources, where we are interested to show not only the relations between data elements within one data source but also the relations between data elements of one data source to data elements of second data source. This is not possible in the standard Chord diagram.

In this paper, we propose an extension to the standard Chord diagram, called **ChordD-2DS (Chord Diagram for 2 Data Sources)**, to show data elements from two heterogeneous data sources and the relations between these data elements. We use a standard Chord diagram for showing data elements and the relations between them from one data source. We arrange data elements of the second data source as an outer layer and show the relations between data elements of both data sources through visual cues. The proposed solution enables us to show data elements of two heterogeneous data sources and their relations in one diagram intuitively rather than multiple diagrams, thus uses space efficiently. We also think that using one intuitive visual diagram reduces load on user working memory rather than working with multiple visual diagrams.

2. The Movie and Twitter Datasets

For a proof of concept in this paper, we use two heterogeneous datasets. The first dataset was obtained using The Movie Database (TMDb) API (<https://developers.themoviedb.org/3>). We collected this dataset for the year 2021 that contains information about a total of 3230 movies, which were further classified as *feature* films

(2417) or *Over-The-Top (OTT)* media service films. Each movie in this category is attached to a number of genres (e.g., action, adventure, animation, comedy, crime, documentary, drama, etc.). In this case, we can see a relationship between different genres based on the common movies between any two genres.

The second dataset is obtained from Twitter, as people’s tweets towards movies can tell us about their sentimental polarity reaction (i.e., *positive*, *neutral*, or *negative*) towards these movies. We collected people’s tweets between the time period of January 01 to September 30 (2021) related to the movies in the first dataset using the Twitter API (<https://developer.twitter.com/en/docs/twitter-api>). We filtered the tweet data based on USA geo-location and English language, which resulted having a total of 3,053,563 tweets. We use the TextBlob library (<https://textblob.readthedocs.io/en/dev/>), a python based text processing library, for tokenizing keywords and to get the sentimental polarity of these tweets.

3. The ChorD-2DS Diagram

In our proposed solution (Figure 1), we use the main Chord diagram for showing data elements and the relations between them from one data source, while we use an outer circular layer for showing the data elements from the second data source. In Figure 1, we use the main Chord diagram for the movies dataset. In this case, we use the inner layer nodes to represent movie genres, where the length of the node represents the number of launched movies in this genre during 2021, e.g., we see more drama and Comedy movies were launched in 2021. We also attached a bar to each inner node (genre) to show the number of attached movies based on *feature* (light brown color) or *OTT* (dark brown color). The code (relation) between any two nodes (genres) represents the number of movies attached to both genres, where the width of the chord represents the frequency.

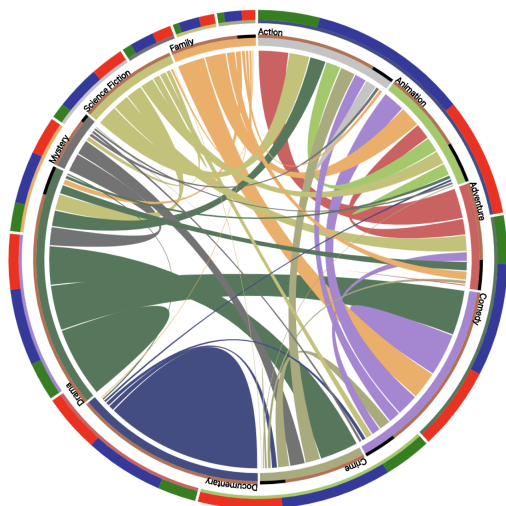


Figure 1: The proposed ChorD-2DS solution, where inner layer nodes represent genres from the movies dataset and outer layer nodes represent genres’ sentimental polarity from the tweet dataset.

For showing the data elements from the second dataset (i.e., the

tweet dataset), we use the outer layer nodes. It is important to note that the total data points in tweet dataset (i.e., total tweets) are not equivalent to the total data points in the movie dataset (i.e., total movies); therefore, length of a node in the inner layer represents different frequency than a node in the outer layer. In the outer layer of Figure 1, we use nodes’ length to show the frequency of associated tweets to each genre and the color inside it to show the overall sentimental polarity (green for *positive*, blue for *neutral* and red for *negative*). The color length shows the frequency of the associated tweets within this genre having this sentimental polarity. To reduce the visual cluttering, we avoided duplicating nodes’ names in both layers. Instead, a thin bar is attached to the nodes of the outer layer with the same color of the corresponding nodes in the inner layer to provide an association between the nodes in the two datasets.

We provide filtering options based on data elements from both datasets, e.g., movies can be filtered based on *feature* films or *OOT* films, tweets can be filtered based on geo-location (e.g., state) or time interval. Mouse hovering over a node at any layer opens a tool-tip with further information about tweets, sentimental polarity, or genres. Also, mouse hovering over any node at any layer fades all other nodes and relations at both layers and keeps highlighting only the same nodes at both layers. For example in Figure 2, we see mouse hovering over the *Documentary* genre in tweets dataset (outer layer) shows a tool-tip with tweets and sentimental polarity details. Further, it fades all other nodes in the outer layer as well as nodes and the relations in the main Chord diagram except the *Documentary* node and the associated relations..

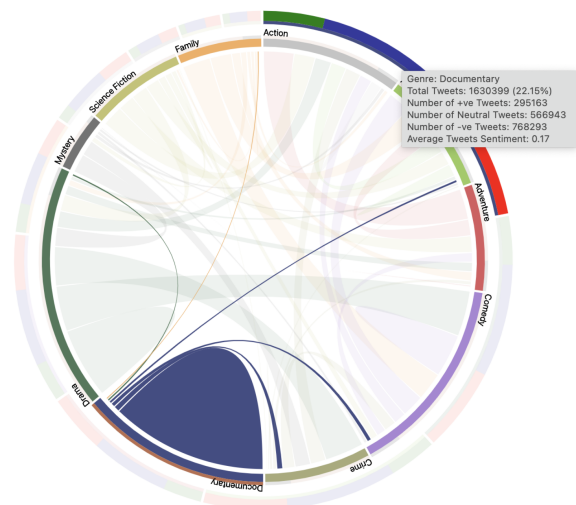


Figure 2: Mouse hovering over a node at any layer fades all other nodes and the non-associated relations.

Our proposed solution is useful in applications where we need to show the relations between people’s feedback on social media (e.g., sentiments in tweet data) with the associated dataset from another source (e.g., COVID-19 dataset or movie dataset), as it provides a compact resulting visualization. In the future, we intend to perform user studies to evaluate whether the proposed diagram reduces load on user working memory compared to using two or more separate diagrams for datasets from two heterogeneous sources.

References

- [HAAE17a] HUMAYOUN S. R., ALI S. H., ALTARAWNEH R., EBERT A.: An Interactive Visual Representation to Explore Association with Hierarchical Social Circles. In *EuroVis 2017 - Posters* (2017), Puig A. P., Isenberg T., (Eds.), The Eurographics Association. doi:10.2312/eurp.20171163. 1
- [HAAE17b] HUMAYOUN S. R., ARDALAN S., ALTARAWNEH R., EBERT A.: TExVis: An Interactive Visual Tool to Explore Twitter Data. In *EuroVis 2017 - Short Papers* (2017), Kozlikova B., Schreck T., Wischgoll T., (Eds.), The Eurographics Association. doi:10.2312/eurovisshort.20171149. 1
- [HBA18] HUMAYOUN S. R., BHAMBRI K., ALTARAWNEH R.: BiD-Chord: An Extended Chord Diagram for Showing Relations between Bi-Categorical Dimensional Data. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces* (New York, NY, USA, 2018), AVI '18, Association for Computing Machinery. URL: <https://doi.org/10.1145/3206505.3206570>, doi:10.1145/3206505.3206570. 1
- [Jal16] JALALI A.: Supporting Social Network Analysis Using Chord Diagram in Process Mining. In *BIR* (2016). 1
- [Krz] KRZYWINSKI M. I.: Benefits of A Circular Layout. *Circos*. URL: http://circos.ca/intro/circular_approach/. 1
- [KSB*09] KRZYWINSKI M., SCHEIN J., BIROL I., CONNORS J., GASCOYNE R., HORSMAN D., JONES S. J., MARRA M. A.: *Circos: An Information Aesthetic for Comparative Genomics*. *Genome research* 19, 9 (2009), 1639–1645. 1
- [QAML17] QI W., ABEL G. J., MUTTARAK R., LIU S.: Circular Visualization of China's Internal Migration Flows 2010–2015. *Environment and Planning A: Economy and Space* 49, 11 (2017), 2432–2436. URL: <https://doi.org/10.1177/0308518X17718375>, arXiv:<https://doi.org/10.1177/0308518X17718375>, doi:10.1177/0308518X17718375. 1
- [TNYT07] THE-NEW-YORK-TIMES: Close-Ups of the Genome, Species by Species by Species. (January 22, 2007). URL: http://www.nytimes.com/imagepages/2007/01/22/science/20070123_SCI_ILLO.html. 1
- [ZSB*17] ZOUN R., SCHALLERT K., BRONESKE D., HEYER R., BENDORF D., SAAKE G.: Interactive Chord Visualization for Metaproteomics. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)* (2017), pp. 79–83. doi:10.1109/DEXA.2017.32. 1