horizontal line



Lyrics Space Odyssey

Design Report

17.6.2020

Group Members

Tsz Kin Chau tszkin.chau@student.kuleuven.be

Fengan Li fengan.li@student.kuleuven.be

Hugo Montaño Castillo hugoenrique.montanocastillo@student.kuleuven.be

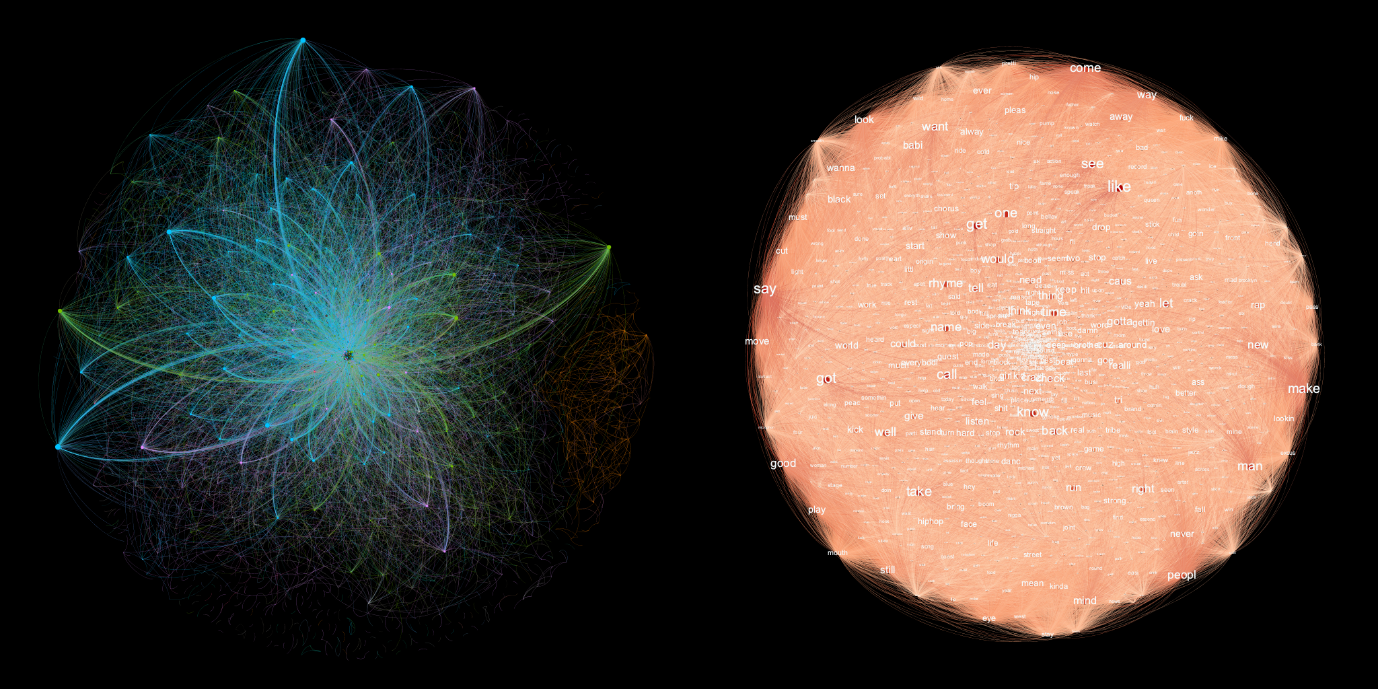
Lars Magne Tungland larsmagne.tungland@student.kuleuven.be

Martina Verna martina.verna@student.kuleuven.be

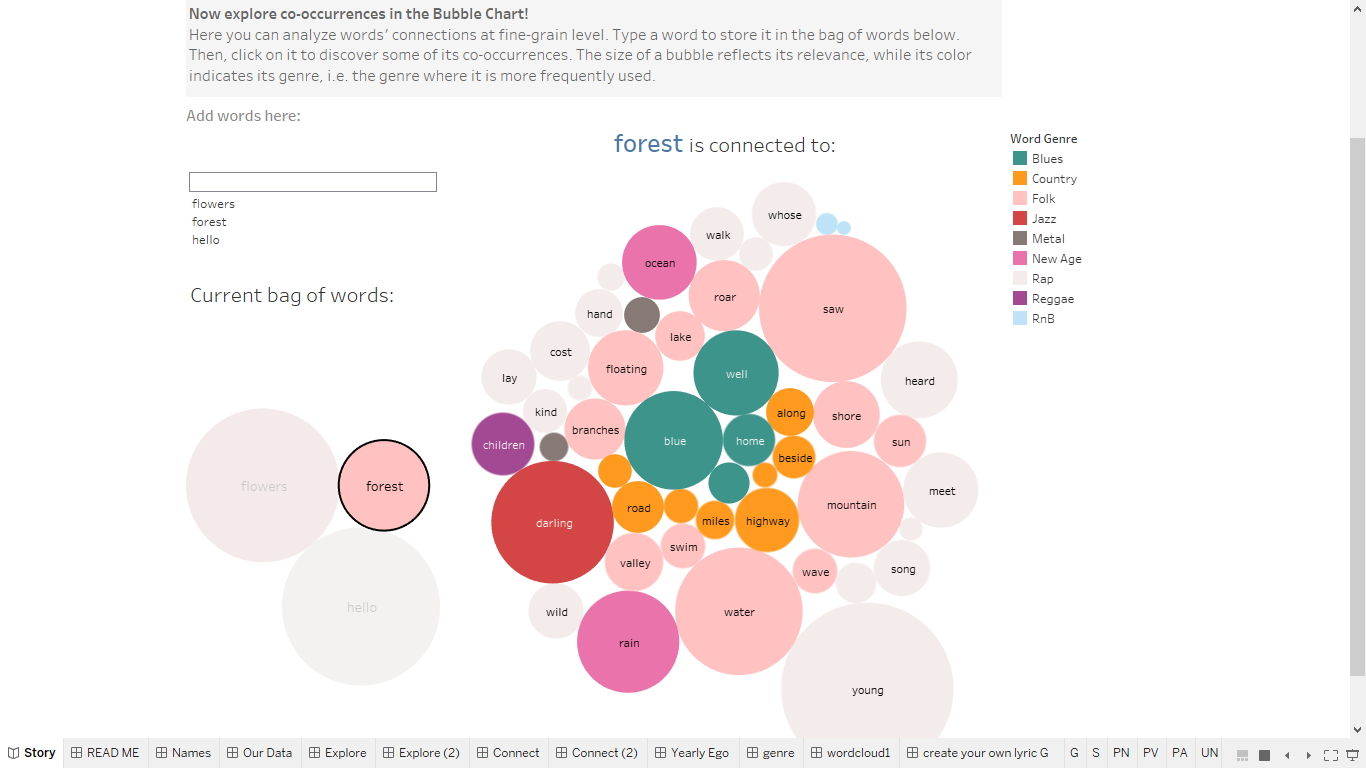
Xin Wang xin.wang@student.kuleuven.be

I. Screencaps

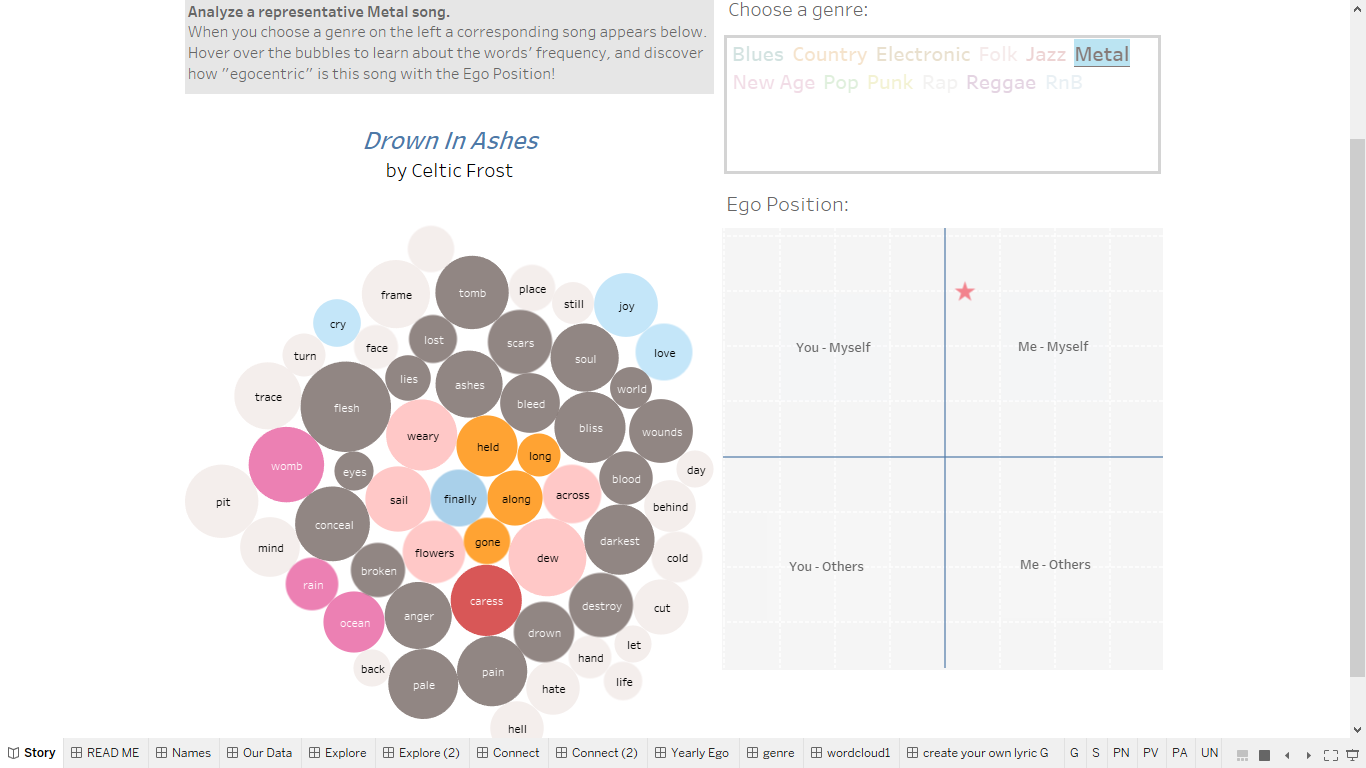
Lyrics and Word Co-occurrence Networks



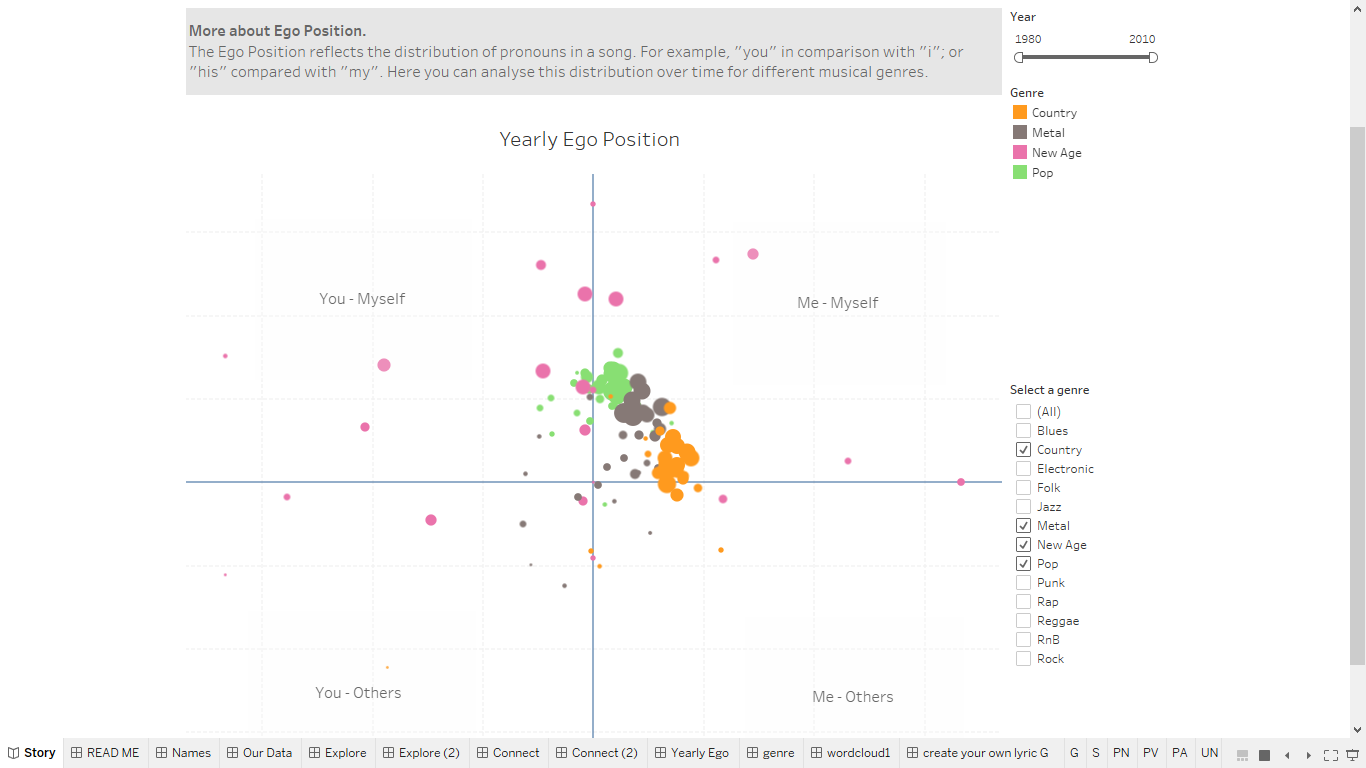
Co-occurrence - Packed Bubbles



Exemplar Song Analysis – Packed Bubbles and Scatter Plot



Yearly Ego Position – Scatter Plot



Packed Bubble Chart – Create Your Own Lyrics

A picture containing screenshot

Description automatically generated

Chord Diagram - Hot Song Titles

A picture containing ride

Description automatically generated

II. Insights

Lyrics and Word Co-occurrence Networks

Vocabularies are shared across music genres with rap having its own.

Co-occurrence Packed Bubbles

Exemplar word co-occurrences and their relative relevance.

Exemplar Song Analysis – Packed Bubbles and Scatter Plot

Which words characterize a song. What is the most frequent genre of this words, and how egocentric is this song.

Yearly Ego Position – Scatter Plot

Evolution of the egocentricity of music genres. Consistency of a genre in the use of pronouns. How this use of pronouns is different among other.

Packed Bubble Chart – Create Your Own Lyrics

Insights are generated by the user, who is given full autonomy to utilize the interactive exploratory tool.

Chord Diagram - Hot Song Titles

Hot titles use uncommon word pairs to create descriptive and memorable titles. Poorly-ranked songs use more common words. We can also use this to create new titles for existing songs.

|  |  |  |
| --- | --- | --- |
| **Artist** | **Original Title** | **Generated Title** |
| Kesha | Tik tok | Rest the Day, Out All Night |
| TLC | No Scrubs | You Ain’t Got It |
| John Denver | Country Roads | Wake Me Up On the Way |

Some common words are not here, most noticeably being the word “Love.” “Love” is one of the most used, or overused words, in every bottom-ranked song. Great love songs describe the theme, they don’t need to put it directly in the title.

III. Data

Lyrics and Word Co-occurrence Networks

In the lyrics network, each node represents a song, whose genre is encoded by color. Each link represents a connection if two songs share over 30% of vocabularies in their lyrics.

In the word co-occurrence network, each node is labelled by the word it represents. Nodes are linked when they appear in the same song.

Co-occurrence Packed Bubbles

In the bag of words there are three variables encoded: the word in question (text), the most frequent music genre for the word (color), and the tf-idf weight of the word (bubble size), the latter considering the whole dataset[[1]](#footnote-1).

In the connected words there are also three variables: connected word (text), most frequent genre (color), and a variation on the tf-idf weight (bubble size). In this case we have the product of the dataset idf, the in-genre idf, and the above average use of a word [[2]](#footnote-2).

Exemplar Song Analysis – Packed Bubbles and Scatter Plot

For each music genre, there is an exemplar song associated. Each music genre is encoded on a specific color (this color also links a word to a genre). And the tittle of the associated song is encoded as tooltip.

The song’s lyrics are encoded as in the Co-occurrence Bubbles chart: the word in question (text), the most frequent music genre for the word (color), and the tf-idf weight of the word (bubble size).

The Ego Chart positions a song in one of the following quadrants: “You-Myself”, “Me-Myself”, “You-Others”, or “Me-Others”. The positioning is done by placing a star in the corresponding quadrant. This reflects the ratio-use of specific pronouns. For example, horizontally we have . While vertically we have something analogous but for the first and second person plurals and possessives, and the third person singular and possessives. Thus, a positioning on a quadrant reflects the dominance of the pronouns. The distance from the axis is a measure of this dominance, the concrete measures are encoded as tooltips.

Yearly Ego Position – Scatter Plot

This is a development of the Ego Chart in the Exemplar Song Analysis. The different quadrants measure the same ratio-use of pronouns, but now for the different music genres. Each genre is encoded in its respective color, each bubble encodes a year, and the size of the bubbles encode the number of songs in that year (number of songs in our dataset). Thus, bigger bubbles indicate more songs and are more representative.

Packed Bubble Chart – Create Your Own Lyrics

TBC

Chord Diagram - Hot Song Titles

Each node represents a word which is used in a song title. Each link between nodes represents a pair of words which are used together in a song title.

The size of each node on the border represents the relatively popularity of the word. The magnitude of the linkage (occurrences of each word pair in all song titles) is shown as a tooltip.

IV. Tasks

1. General design concept: Describe, Explore, Create.
2. Lyrics and word co-occurrence networks task: Contextualize the data.
3. Co-occurrence Bubbles: explore a word’s connections with other words.
4. Exemplar Song Analysis: explore the characteristics of different songs.
5. Yearly Ego Position: explore the changes in the use of pronouns.
6. Packed Bubble Chart Task: create song lyrics using words from the Million Songs Dataset.
7. Chord Main Task: create new and interesting song titles.

GENERAL DESIGN CONCEPT

Lyrics Space Odyssey is a choral project composed of a series of visualizations, which all together contribute to displaying different – but complementary - aspects of the same data.

From a broader point of view, the choice of using specific visualization tools must be considered in itself a design choice, which we conceived as functional to the development of a precise data story. In fact, our visualizations can be divided into 3 sets (network visualizations, bubble charts and chord diagrams), corresponding to the 3 different tools we used (Gephi, Tableau, D3 visualizations).

The final narrative, which resulted in a Tableau Story presentation, is also divided into 3 conceptual "chapters": **Describe**, **Explore**, **Create**. Ideally, the sequence of the dashboards in our storytelling reflects our aim to take the user through a journey into song lyrics, starting from a more general level down to individual songs and lyrics.

In this sense, the Gephi networks present an introductory overview, a real “galaxy” of edges and nodes: the user here detects color clusters and perceives the complexity of the data, so that he wants to know more about them; in the second section, a set of interactive bubble charts and scatterplots allows the user to explore the data more in depth, as well as to get his own insights while performing simple actions (select and focus, filter, connect); finally, in the last section, the chord diagrams are meant to engage the user in a creative challenge consisting in the identification of “hottest” and “less hot” features of song titles.

Lyrics and Word Co-occurrence Networks

1. Lyrics network: Encode the relevant connections of lyrics among songs and genres.
2. Word co-occurrence network: Map the pattern of co-occurrent words in different song.

Co-occurrence Bubbles

1. Explore the preset co-occurrences.
2. Type more words and explore its co-occurrences.

Exemplar Song Analysis – Packed Bubbles and Scatter Plot

1. Choose a genre and compare exemplar song lyrics.

Yearly Ego Position – Scatter Plot

1. Explore changes in egocentricity (epochs and genres) by adjusting the filters.

Packed Bubble Chart – Create Your Own Lyrics

1. Allow and encourage the user to create lyrics using words from the Million Songs Dataset
2. Explore word connections interactively and uninterruptedly
3. Gain insight about the majority genre, popularity and uniqueness of the connected words

Chord Diagram - Hot Song Titles

1. Allow user to create new and interesting song titles
2. Highlight differences in word usage between the hottest and least hot songs

V. Design Decisions: Visual Encodings & Interactions

Lyrics and Word Co-occurrence Networks

Principle:

To visualize interesting patterns without predefined ideas and biases.

Iteration:

The very initial design idea germinated from the curiosity about potentially interesting patterns in the data. Therefore, subsequent effort was mainly devoted to explore the implicit connections contained in the dataset and to visualize them via networks. As the first attempt, the word co-occurrence network visualizes the connections between words, which appeared less than satisfactory. A second attempt was then made to construct a lyrics network in the light of genres, which turned out to show more interesting insight: rap stands out and shows a rather unique vocabulary field.

Encoding nodes and edges:

The most crucial design decision was to set the cut-off point of overlapping song vocabularies to 30%, which was proved to present the best visual appeal. In other words, songs that share over 30% of identical words in their lyrics are linked, while the rest gets excluded from the network. It is to be noted that the word-list dataset contains merely 5,000 words, among which English words amount to less than 3,000. This relatively small bag of words inevitably causes certain bias, as words that do not appear in this list are not considered, which diminishes the uniqueness of each music genre.

Color:

As the edges are thin in shape and large in quantity, the focus for color choice is centered on optimizing the visibility of all links. A dark background was chosen to foreground the bright-colored edges with high value and high saturation. This way, the thin lines first of all stand out from the background, while genres can also be clearly differentiated from each other thanks to the high contrast. As a pleasant byproduct, the final graphs create a gleaming and cloudy effect.

On the other hand, while color plays a role in the lyrics network, it does not have a function in the word co-occurrence network.

Co-occurrence Bubbles

Principles:

The main principle behind this visualization is to display information at a glance. The encodings and concepts used all along chapter 2 are introduced here. Thus, this visualization aims at **precise and intuitive** presentation of the data.

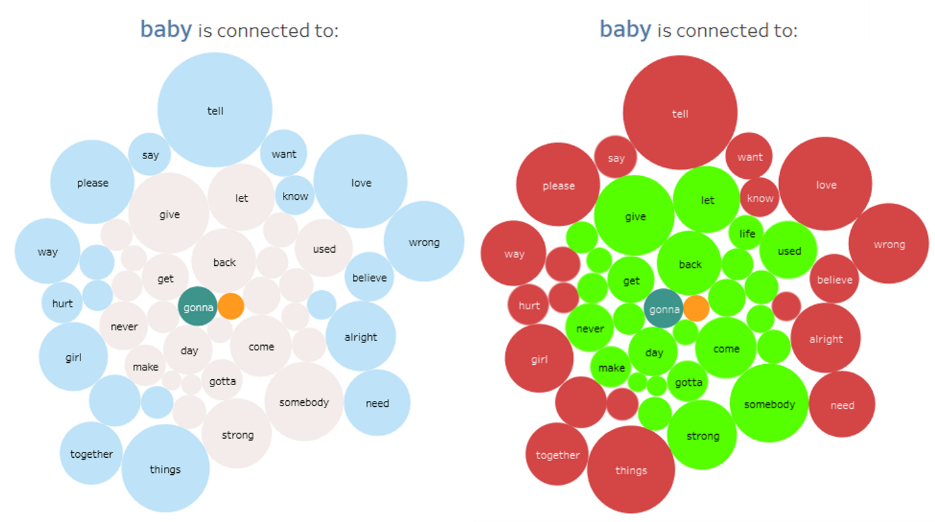
The variables/concepts introduced are music genres, co-occurrences, and the relevance of these co-occurrences. A Packed Bubbles chart is used for encoding them.

General Display:

This dashboard consists of two packed bubbles charts. On the left, there is a bag of words (BOW) where the words-to-explore are kept. On the left, there is a second packed bubbles chart which shows the co-occurrences of a selected word in the BOW. The reason for having two charts was to separate the actions that a user could take. On the one hand, the collection and selection of words is kept in the BOW. On the other hand, the exploration of the co-occurrences is kept on the right. Both charts have the same encodings. A series of instructions on how to interact with the dashboard are provided.

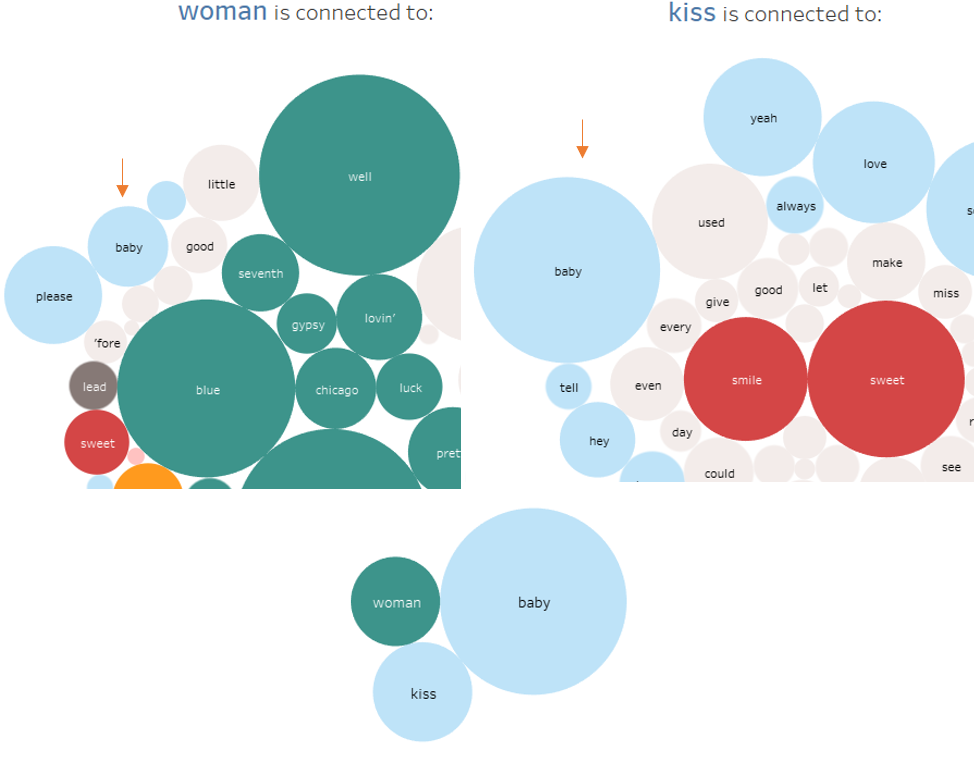
Colors:

The categorization of the words is encoded with different colors. The color palette responds to the distribution of words. Our analysis revealed that Rap was the genre with the biggest number of words, followed by R&B. And most importantly, these words were often used in all songs. Thus, Rap and R&B are encoded with low saturated colors: light grey and light blue, respectively. The opposite was done for the genres with a small number of words, for instance Blues and Jazz. These genres were encoded with highly saturated colors. The objective was to avoid scenarios like the one in the figure below, where Rap and R&B could lead to big areas of saturated color.



Bubble Size:

The size of the bubbles indicates the relevance of a word. In the BOW the size reflects the td-idf weight of the selected word considering the whole dataset. In the co-occurrence chart the size also reflects the idf weight the genre of the word selected on the BOW. Take the example in the figure below. In this case “baby” (an R&B word) is retrieved as co-occurrence of “woman” and “kiss” (Blues and R&N words, respectively). The different sizes for “baby” are due to the different genres of the words. “woman” is a Blues word so its connection to the R&N word “baby” is less relevant. On the contrary, “kiss” is also an R&B word so its connection to “baby” is more relevant. Nevertheless, “baby” is a quite common word, so its size is comparable to that of less common words like “blue” or “sweet”.



Interaction:

The main feature of this visualization is the display and retrieval of relevant co-occurrences. By clicking on a word in the BOW, its connections are displayed on the right side of the dashboard. If a user would like to explore other words than those in the presetting, she can do so. For this purpose, there is a dedicated box where she can type a word of her interest.

Exemplar Song Analysis – Packed Bubbles and Scatter Plot

Principles:

This dashboard also follows the purpose of providing information at a glance. But now, regarding a song’s lyrics. In addition, a new concept is introduced: the egocentricity of a song.

General Display:

This dashboard is also built around two parts. On the left there is a packed bubbles chart like the one detailed in the previous section. The novelty is the use of a scatter plot to measure the egocentricity of a song.

Packed Bubble chart:

This chart represents the co-occurrences of words within a song. It keeps the conventions introduced in the previous packed bubbles charts: color for word-genre and size for words’ relevance. The difference relies on the setting of the relevance. Because this chart displays words from a specific song, we are dealing with **co-occurrences within a song**. In this vein, a bubble’s size reflects the relevance of the word within the song. It corresponds to the simple tf-idf weight of the word: the idf for the whole dataset and the tf as a term’s usage in the song.

Ego Scatter Plot:

A scatter plot is an excellent way to compare bivariate data. This visualization allows the precise representation of an otherwise too abstract variable: pronoun usage. In our visualization each axis encodes a dimension of egocentrism. Horizontally, the relation between the first (I) and second (You) person singular. Vertically, the first person possessive, object pronoun and its plurals (me, my, we, our, us) compared with the second person possessive and third person singulars, plurals, possessives and object pronouns (your, he, his, him, she, her, they, their, them) . The more dominant a pronoun (or pronouns for the vertical axis), the further placement from the center.

The main aspect of this scatter plot is the position and distance from the center, but to enhance the visualization we adjusted some characteristics. **Color**: to improve the contrast with the dashboard’s background, we gave the scatter plot’s background a light grey color. To divide the quadrants, the axis are highlighted with a saturated color. The position of a song uses a shape and color that contrast with the rest of the dashboard. **Labels**: since the focus of the chart is the position and not the actual measures, we suppressed all the axis labels and marks. But we labeled the different quadrants with a legend box for a better identification and understanding of a quadrant. **Grid lines**: white grid lines were kept to provide a positional reference.

Interaction:

To allow the exploration of the different exemplar songs, there is a box with the different music genres. This box simply displays the names of the genres in its corresponding color. The user may click on any of these genres thus, triggering the display of a song. Initially, the user had the possibility of searching a song of her interest, but this could lead to frustration after repeatedly failing to find a specific song. So, we restricted the selection to particular songs.

Yearly Ego Position – Scatter Plot

Principles:

This dashboard emphasizes the aim of giving insights at a glance. This time the main objective is to contrast the egocentricity of the different music genres.

General Display:

This dashboard is an elaborated version on the ego chart described before. Instead of positioning a unique song, in this chart the user discovers the ego positioning of a music genre as a whole. For this purpose, more variables were added to the basic ego chart: music genre, year, and number of songs for each year. To control the display of data, some filters were added: one for the genre and one for the years.

Shape:

To keep the theme in the exploratory chapter, the shapes in this scatter plot are circles. This capitalizes on the pre-attentive processing from the previous visualizations. The user remains clustering circles of same color. But in contrast to the previous charts, now she finds a circle for each year-genre combination.

Color:

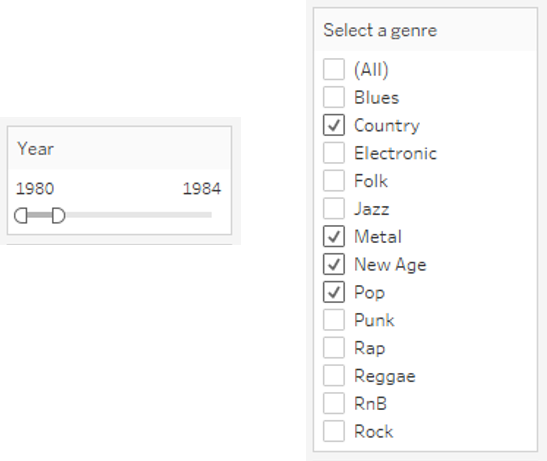
Since each circle corresponds to a music genre the same color palette is kept.

Size:

The mayor novelty in this chart is the association of the circles’ size to the number of songs. Our dataset includes a variable number of songs for each year and for each genre. There are years when a genre might have few songs and other years with more songs. We exploited this fact by associating the number of songs with the size of the circles. This functions as an indicator of a year’s relevance: years with more songs are more representative and bigger, and thus, more relevant.

Interaction:

There are two filters for the exploration of the different genres. The filter controlling the years is a sliding filter, this allows for an easy selection of different ranges of years. For controlling the display of specific genres there is a selectable list. This allows a controlled selection of the genres; the user can easily see which genres are currently displayed and which ones are omitted.



Packed Bubble Chart – Create Your Own Lyrics

Principles:

The guiding principle is to find a balance between level of details and cognitive burden to a user. Being an exploration environment as well as a lyrics creation application, the learning curve is inevitably steep. New knowledge a user needs to acquire when using the tool includes, parts of speech (POS), song genre, popularity-uniqueness, and the connection metaphor. Thus, a considerable effort has been devoted to simplifying UI/UX to encourage the user to ‘explore and create’.

Component of the visualization:

1. Mega Bubble Chart: Located on the left of the dashboard, it renders all the word appeared in the Million Songs Dataset. User can see the detail of a word bubble via tooltip and choose a word of interest by clicking it.
2. Connection Bubble Chart: Located on the right of the dashboard, it renders the chosen word and its connected words.

Colors:

Coloring of the nodes represent the major genre of a word. Color scheme follows the convention in the previous visualization.

Background is tuned to medium gray to support nodes with very high or very low brightness value so no nodes will be receded to the background.

Number of nodes:

The control of number of nodes has been one of the most difficult tasks in building the bubble chart. To be an effective tool for lyrics creation, at least a user needs access to noun, verb and adjective / adverb. Without POS division, users have a high probability to get noun connection because the dataset is biased towards noun. Also, in the dynamic of lyric composition, a user needs to pick the right POS to construct verses.

Moreover, we also have to leverage popular words (high term frequency) and unique words (high inverse document frequency). There are frequently appeared words such as ‘love’ and ‘baby’ and unique connections such as ‘forest’ -> ‘enchanted’, ‘spiritual’.

For the first challenge, we have to find a way to separate POS so a user may focus on the designated POS. For the second challenge, we have to find a way to return words from both the popular side and the unique side. As color is reserved for majority genre and size for popularity, virtually there is no further mean of encoding available in the Tableau packed bubble chart.

Initial attempt is to partition the bubble chart into 6 parts, representing popular noun, verb, adjective/adverb and unique noun, verb, adjective/adverb respectively.

A picture containing screenshot

Description automatically generated

This layout has a few issues. First, it is difficult to find a focus among six partitions and this merely increases the cognitive burden of a user. Second, because Tableau will size the bubbles according to the active data in the view, thus, even with lower frequency, unique words have a same scale as the popular words.

Much effort is invested in suppressing the excessive details in the fragmented six partitions layout. By combining top popular connections and top unique connections with top\_n\_filter, and by putting up a dropdown menu for POS option. Finally, a single bubble chart is layout is made possible.

top\_n\_filter is defined as,

IF ATTR([word\_genre\_label\_mask]) = '\_self' THEN "your selected" ELSE

IF [rank\_top] <= 20 - ([connection\_threshold]-1)\*10 THEN "a popular" ELSE

IF [rank\_bottom] <= ([connection\_threshold]-1)\*10 THEN "a unique" ELSE NULL END END END (?)

The final layout contains the node count of 1 chosen word plus 20 connected words.

Size of nodes:

The size of the nodes is proportional to the harmonized word count. In Create Your Own Lyrics, node size is encoded to express the popularity (frequency) of a word which in turn allows the user to learn the popularity difference between the chosen word and the connected words. The challenge here is that the returned 20 connected words are derived from two different ranking, top term frequency and top inverse document frequency. Therefore, a metric common to both ends is necessary to interface the sizing.

A number of metrics are tested for sizing. Namely,

MAX([Count])

AVG([Count])

MAX([termFrequency])

MAX([Count])/AVG([Count])

sqrt(SUM([Count]))

|  |  |  |
| --- | --- | --- |
|  | A close up of a logo  Description automatically generated |  |
| MAX([termFrequency]) | sqrt(SUM([Count])) | AVG([Count]) |
| Comparison between three metrics. Note that in navy blue the chosen word has 616 occurrences while the connected words have occurrences between 19 and 16,869. | | |

In the end, sqrt(SUM([Count])) is picked as the harmonized word count.

Relation between nodes:

The bubble chart is a representation of a ‘origin-edge’ relation in which the chosen word is the departure point (origin) and the connected words are edges. The relation is realized by always positioning the chosen word (colored in navy blue) at the center.

|  |  |
| --- | --- |
| page159image61378000 | A close up of a logo  Description automatically generated |

Interaction: Wildcard Filtering

User is allowed to search a word from the mega bubble chart to explore word connection. Wildcard filtering is used in which a user may feel that he/she is progressively navigating in the word bubbles. This is to immerse the user in the meta-narrative of ‘Lyrics Space Odyssey’.

|  |  |  |
| --- | --- | --- |
| A picture containing food  Description automatically generated | A close up of a logo  Description automatically generated | A picture containing device  Description automatically generated |
| search string: empty | search string: ‘for’ | search string: ‘forest’ |

Interaction: Exploring Word Connection

Clicking any word bubble in the dashboard will trigger the connection bubble chart to display its word connections. To encourage a user to explore word connections, the interaction is designed to be as fluid as possible. The main issue is that it is easy for a user to lose his/her whereabout after exploring a few rounds: which word am I exploring?

A picture containing screenshot

Description automatically generated

The initial attempt is to use a word bubble dedicated as the indictor of the chosen word. During our internal review, we discovered that it is not intuitive for people to recognize the dedicated bubble as the chosen word. Then, we tried to modify it by adding a scenario sentence ‘connection to <word> in <genre> songs’ to provide contextual information.

This could not solve the problem as well. First, we think that the scenario sentence adds complication to the UI. Second, both the dedicated word bubble and the scenario sentence inevitably cause the user to move their eyeball back and forth. We consider these an increase of cognitive burden and a hindrance to the fluidity of the interaction.

|  |  |
| --- | --- |
|  | A picture containing stereo, sitting, display, holding  Description automatically generated |

Then, we try to solve it by embedding ‘my whereabout’ within the bubble chart itself.

We tried to make use of pre-attentive processing of the brain. First, we added the chosen word in the connection bubble chart and colored it in navy blue. It is based on the fact that our brain can quickly detect object with distinctive coloring. Second, we instructed Tableau to always position the chosen word at the center. It is based on the cultural phenomenon that centrality tends to have a special status. Third, we made use of the highlighting function. Once a user selects another word to trigger the re-render of the connection bubble chart, the newly chosen word will be highlighted.

We discovered that our brain can quickly learn the paradigm and the connection bubble chart is made self-explanatory. The value of this more fluid version is that it encourages the user to

make more connection exploration. This aligns will our meta-narrative of ‘lyric space odyssey’.

Interaction: Panning Popularity-Uniqueness

User is allowed to pan across the Popularity-Uniqueness spectrum with a toggle. Learn may learn the relative popularity or uniqueness of the chosen word by comparing with its connected word.

|  |  |  |  |
| --- | --- | --- | --- |
| A picture containing photo, sitting, table, white  Description automatically generated |  | Full-Popular |  |
|  | Mixed |  |
|  | Full-Unique |  |

Other Interactions: Lyric Composer, Genre Switch

User can use the lyric composer to compose lyric. Action available includes add word, backspace and reset. Lyric composer is a Node.js app embed as a web object in Tableau Dashboard. Word collected in the lyric composer can be retrieved again as a data source of Tableau. Additional analysis of the created lyric is possible though it is not implemented in the scope of this project.

Genre switch allow user to navigate to the word connections of songs from another genre. By default, clicking a word bubble will show the word connections in the reprehensive genre of the chosen word. Again, color scheme of the genre follows the convention in the previous visualization.

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated |  |
| Lyric Composer | Genre Switch |

Chord Diagram - Hot Song Titles

Principles:

The driving principle of this chord diagram is ease of interaction. Depending on the luck and interest of the user, they may or may not find a spectacular new title. However, they should feel as if they could and understand how to do so if they spent more time with it.

Every element starts out presented with equal prominence. Admittedly this is somewhat overwhelming at first. However, as soon as the cursor moves over any element, all other elements are immediately suppressed and the mouseover element remains highlighted. Aided by the dark background, the highlighted chords give the user a pathfinding mechanism enabling exploration. In this way, I help align the presentation of the data to the user’s visual focus.

Colors:

Background: I used the RGB value [50, 50, 50]. A dark background greatly enhances readability of fine, light-colored lines. I found that a pure black background creates slightly more uncomfortable contrast against the brightly color palette.

Text: Wheat (#F5DEB3) gives a pleasant contrast to the dark background and does not replicate any of the colors of the chords. Thus, it avoids creating confusion about labels being associated with more than their individual node.

Node and Edge: For convenience, I chose to use D3 palette [Categorical.Set1]. For a diagram of ~30 nodes, each color is repeated about 3 times. To mitigate the impact for those who are red-green colorblind, the two colors are never placed adjacent to each other. Furthermore, bright edges against dark background would still help with finding connections.

Fonts:

Size: Given the density of data, fonts are sized to maximum that will fit on screen. This ultimately became dependent on maximum word length. Height is the limiting factor for desktop displays since they are typically 16:9 aspect ratio. However, width is the limiting factor for portrait mode mobile displays. For example, the Samsung Galaxy S10 is 3200x1440 and the iPhone 11 is 1792x828; both are 9:20 aspect ratio. Theoretically, you can encourage users to turn their phones but it’s an extra second of work which is a meaningful amount. Alas, the old 4:3 aspect ratio is best for circular diagrams but CRTs are as extinct as floppy disks.

Font selected: Arial is the classic and well-suited for our task. It’s web-safe, free and readable. Looking back on it, the kerning is a bit smaller than ideal. Given the odd angles in which text appears, more generous spacing would help with readability even at the expense of screen real estate.

Rotation:

This was a difficult decision. Horizontal text labels improved label legibility but diminished the relationship between labels and nodes. Rotating text labels improved the natural flow of the eye from label to node to chord. However, labels at the very top and bottom of the diagram are difficult to read. Ultimately, I chose to prioritize maintaining the circular theme of the chart. To aid the user, all tooltips are displayed horizontally.

Layout (in the Tableau Story):

Since chord diagrams are not as well-known or self-explanatory, I chose to start by giving the user some samples of what they expect to find. Even if the user does not explore the graph themselves, they don’t leave the slide empty-handed.

Nonetheless, I hope that the large brightly colored diagram in the middle of the page would encourage users to play with it for at least a bit. Please note that the for the sake of time and consistency during the presentation, we replaced the live chart with an animated gif showing how to create the title “Rest All Day, Out All Night.”

Desktop/Mobile:

Because of the number of nodes, this diagram is best suited for desktop usage. While it does dynamically scale, zooming out becomes problematic as discussed in below in “Nodes and Edges”. Mobile also has no concept of cursor which makes the on hover effect much less useful. The alternative using Holoviews is far better suited to mobile use.

Number of nodes and edges:

This was a huge part of testing and probably the most difficult design decision. The more nodes and edges, the more complex and interesting connections are revealed. However, there is a massive tradeoff in usability and readability beyond a certain point. A chart with fewer than 20 nodes dies a rather dull death; there are simply not many interesting titles.

Chaos springs from charts with more than 100 nodes. Nodes are difficult to pick and the thin lines are difficult to follow. Each edge node has an average angular size of 25 pixels. In practice, less popular words will have an edge length of <10 pixels. Compare this the average cell height in Excel of ~40 pixels – I would be limiting the usability of the chart to Excel wizards and competitive FPS gamers.

Size of nodes and edges:

The size of each node is proportional to its term frequency. This creates a problem where common words (e.g. “the”, “you”) take up a huge amount of screen space, thus squeezing other words. Less common, but more interesting, words then become difficult to select. To resolve this, I manually adjusted the raw frequency using roughly a log scale. What this detracts from scientific accuracy, it enhances in usability.

I also implemented a cut-off so that words with fewer than 4 total connections are excluded from the final output. These terminal words add little to the variety that a user may find.

References:

My implementation and customization use the D3 wrapper by https://shahinrostami.com/ and package from http://holoviews.org/.

VI. Alternatives

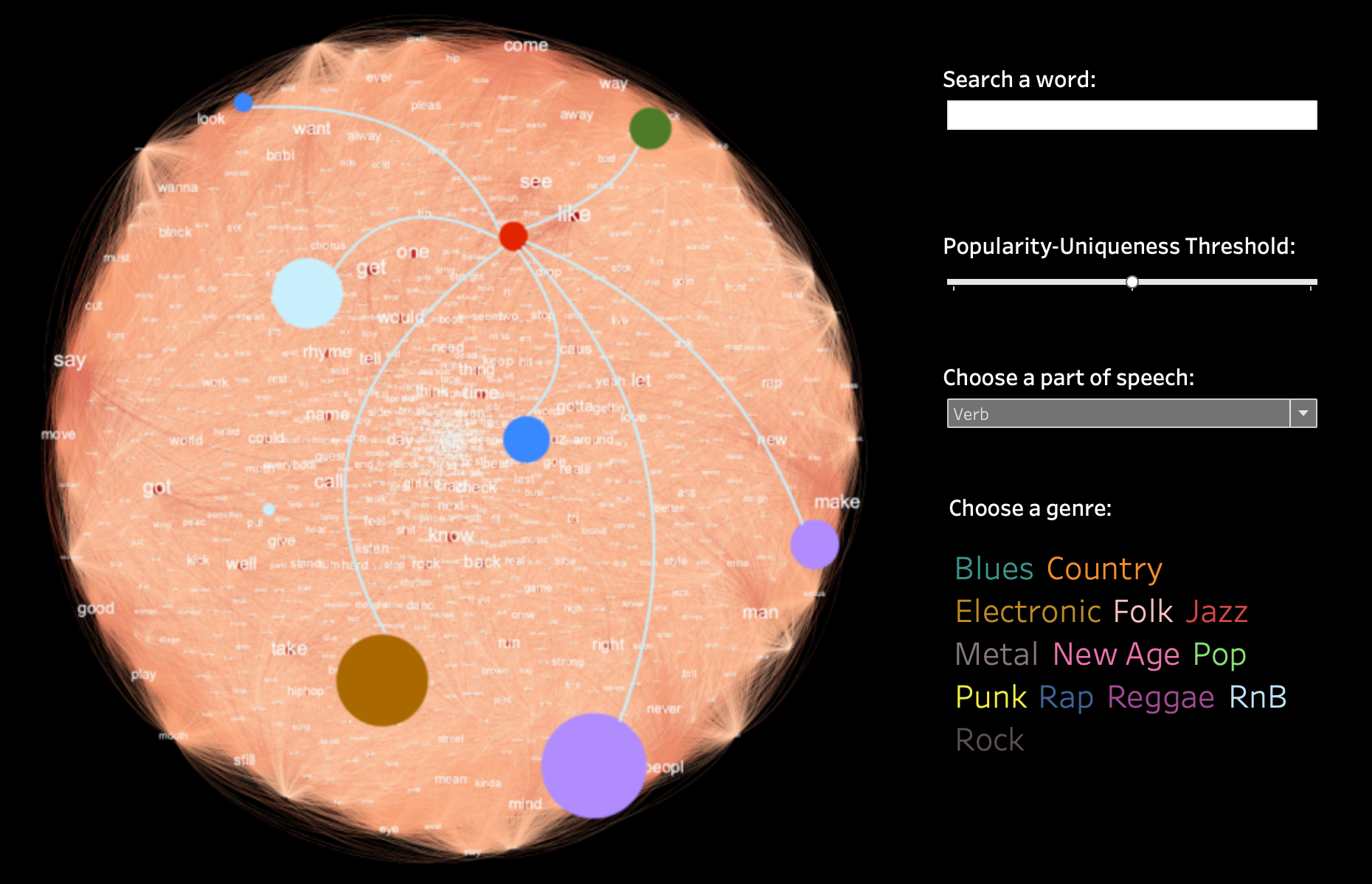
Alt 1 - Word Frequency Cloud

The word co-occurrence network could be a word frequency word cloud. By including the frequency of word occurrence and co-occurance, each word is assigned with a certain weight, which can be visualised for instance in size. This new parameter could create new connections and bring forward different insights. The reason why the current version does not include this aspect is due to its priority to identify unique words, which then made word frequency less relevant.

Alt 2 - Full Network with Enlargement: a Combination of Word Co-occurrence Network and Create your Own Lyrics

Design Decisions

The Mega Bubble Chart and Connection Bubble Chart could be combined and rendered on the word co-occurrence network graph. In other words, the interactive features for user exploration, namely filtering the position of speech, popularity-uniqueness and song genres, could be relocated to the word co-occurrence network graph from bubbles.

To activate the interaction, the user clicks on the labelled nodes in the network, which triggers to magnify and foreground the chosen word’s connections as in the Connection Bubble Chart. Three filter bars are retained on one side to support further exploration.

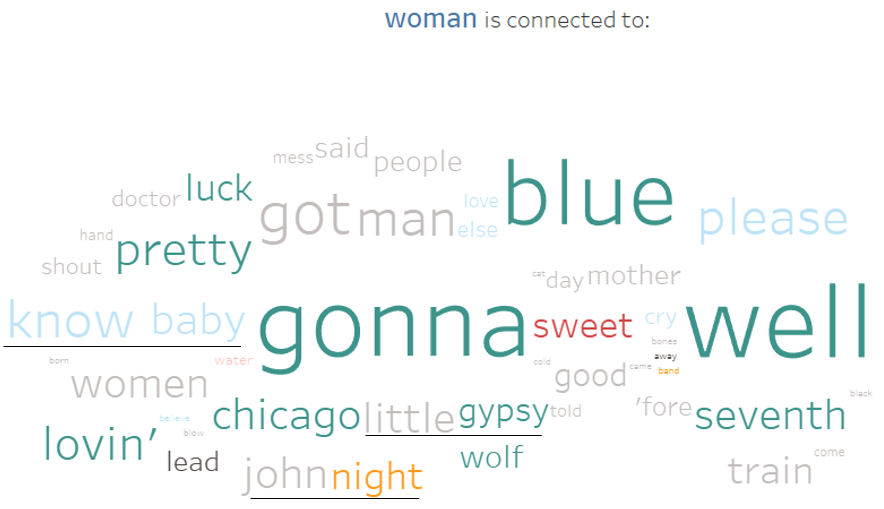
Comparison

The major advantage of this combination is to provide the user with a constantly adjusting contextualization of the data during his exploration. It is as well economic in using space.

Alt 3 - Add Additional Navigational Cues to Create your Own Lyrics (TBC)

Alt 4 - Co-occurrence Word Cloud

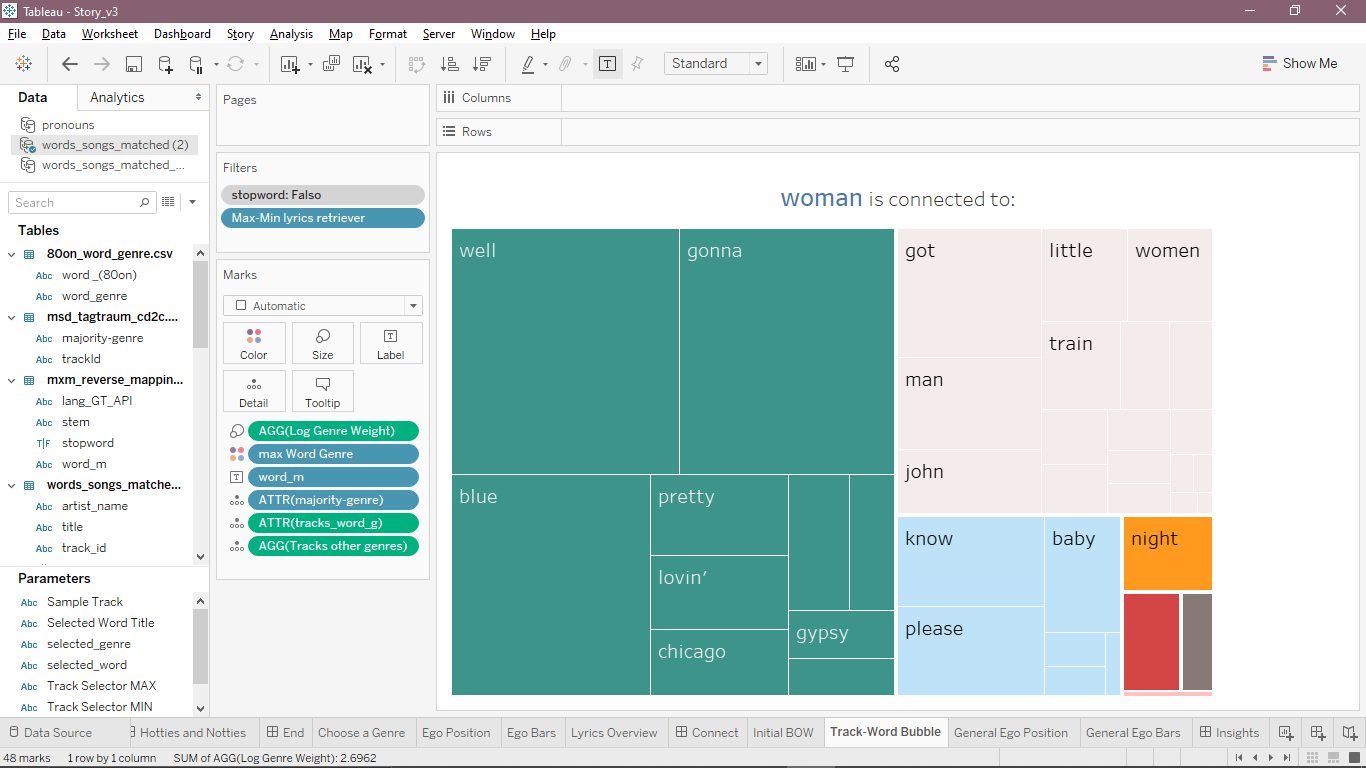
A recurrent way to visualize text is using word clouds. At first sight, one may think that this visualization displays text in a clearer way but, actually, this only holds for the biggest text. In the figure below, one can see cases where the text is so small that only a color line can be distinguished. For our purpose of providing insight at a glance, this visualization falls short. One can grasp the text and nothing else. We wanted to highlight the genre and relevance of a word, but in this case the words’ size could prevent the adequate recognition of the color. Perhaps, another color palette could ameliorate the situation, but then, we could be facing excessive use of color.



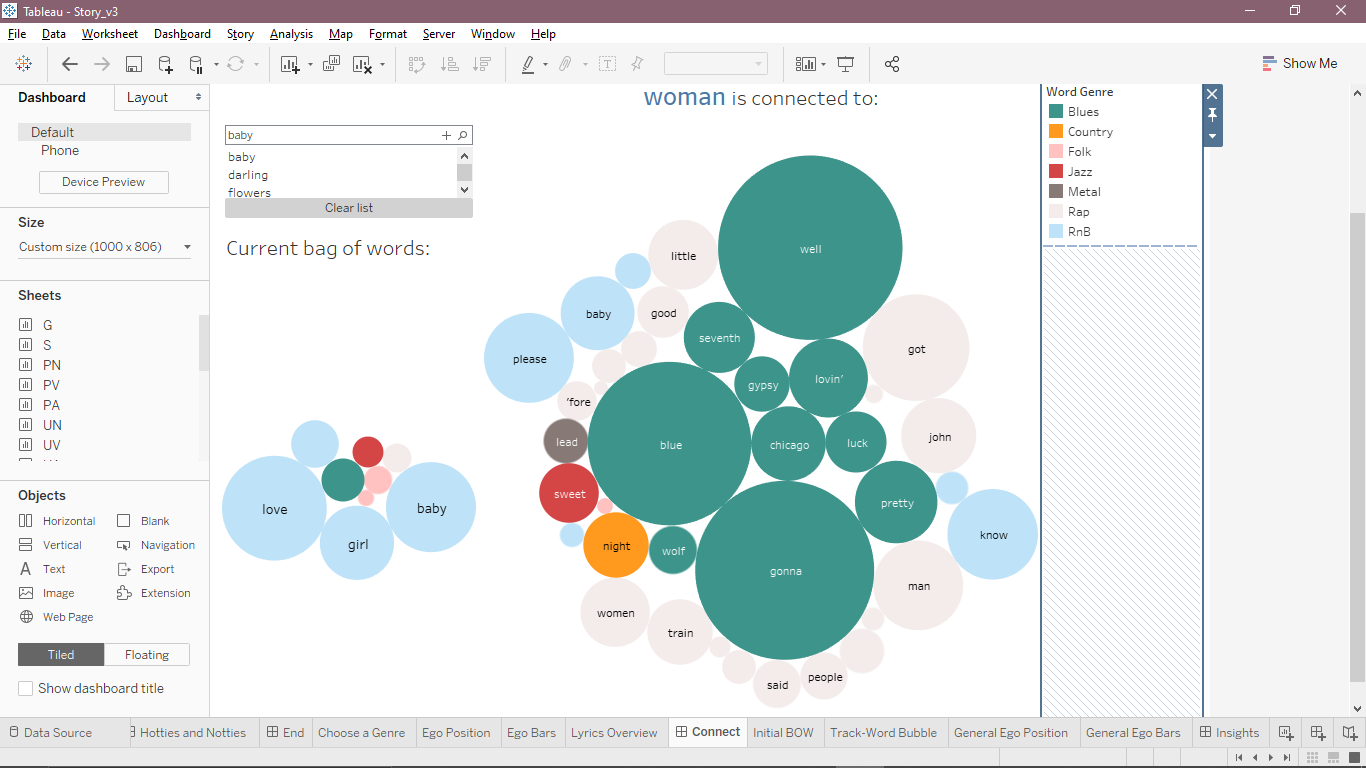
Another issue regarding the word cloud is the lack of separation among words. In the above figure, the words “know” and “baby” have the same color, similar size and are too close. In the hypothetical case of having the words “butter” and “fly”, one could mistakenly read “butterfly”. The closeness among words can also deviate the attention of the user towards the formation of phrases, e.g. “little gypsy” or “john night”.

Alt 5 - Co-occurrence Treemap

Another alternative for visualizing the words co-occurrences could be a treemap. This visualization has the advantage of having a wide area for displaying color, while keeping a sharp separation among figures.

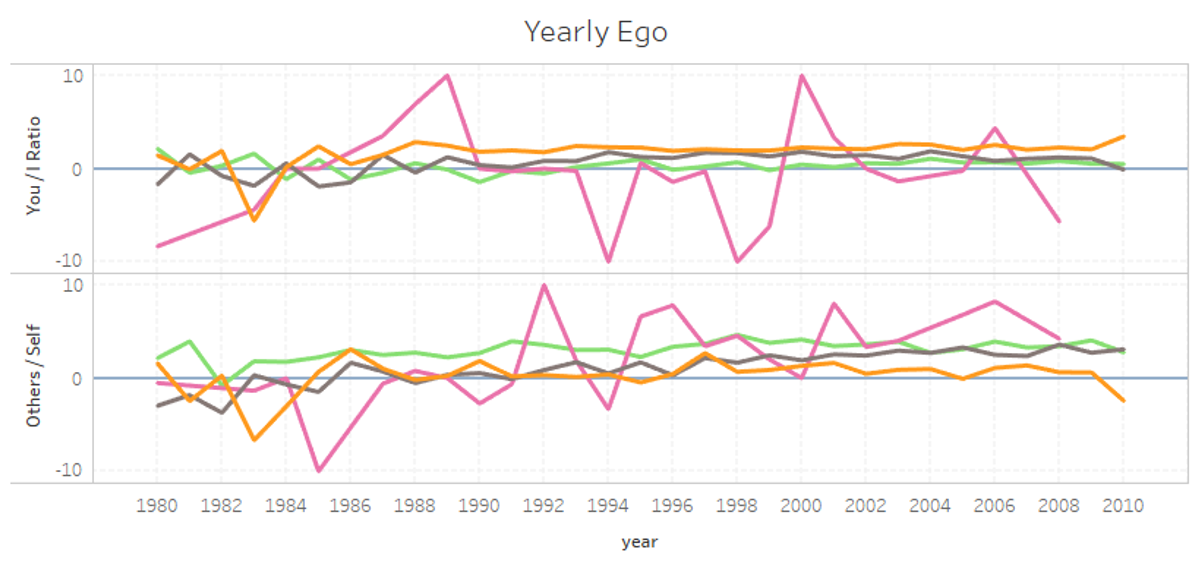


Indeed, this could be an adequate alternative to our packed bubbles chart. But the reason for choosing the bubbles is their aesthetics. We tend to prefer curvy figures rather than straight and edgy ones. While both the packed bubbles and the treemap provide a good overview of the co-occurrences (genre and relevance), the former seem more playful and engaging. Also, the bubbles direct the eye towards the biggest bubbles (the most relevant), while the hierarchical organization of the treemap directs the eye towards the bigger cluster, which might not contain the most relevant words.



Alt 6 - Yearly Ego – Time Series

An alternative encoding for the data of the Yearly Ego chart could be using time series. This visualization provides a detailed overview of evolution of a variable. In this case we could have two parallel series detailing the yearly changes of the scatter plot axis. This kind of visualization is good for displaying the trends of a variable. In the figure below we can see a steady evolution of the orange, green, and brown lines, while the pink line shows a crisscross pattern. The main issue with this visualization is the impossibility of encoding more variables in a compact way. In the time series the variables have to be separated, and by doing this we lose the possibility of encoding the number of songs in an effective way. The number of songs could be added on an additional axis, which leads to an even more complicated visualization.



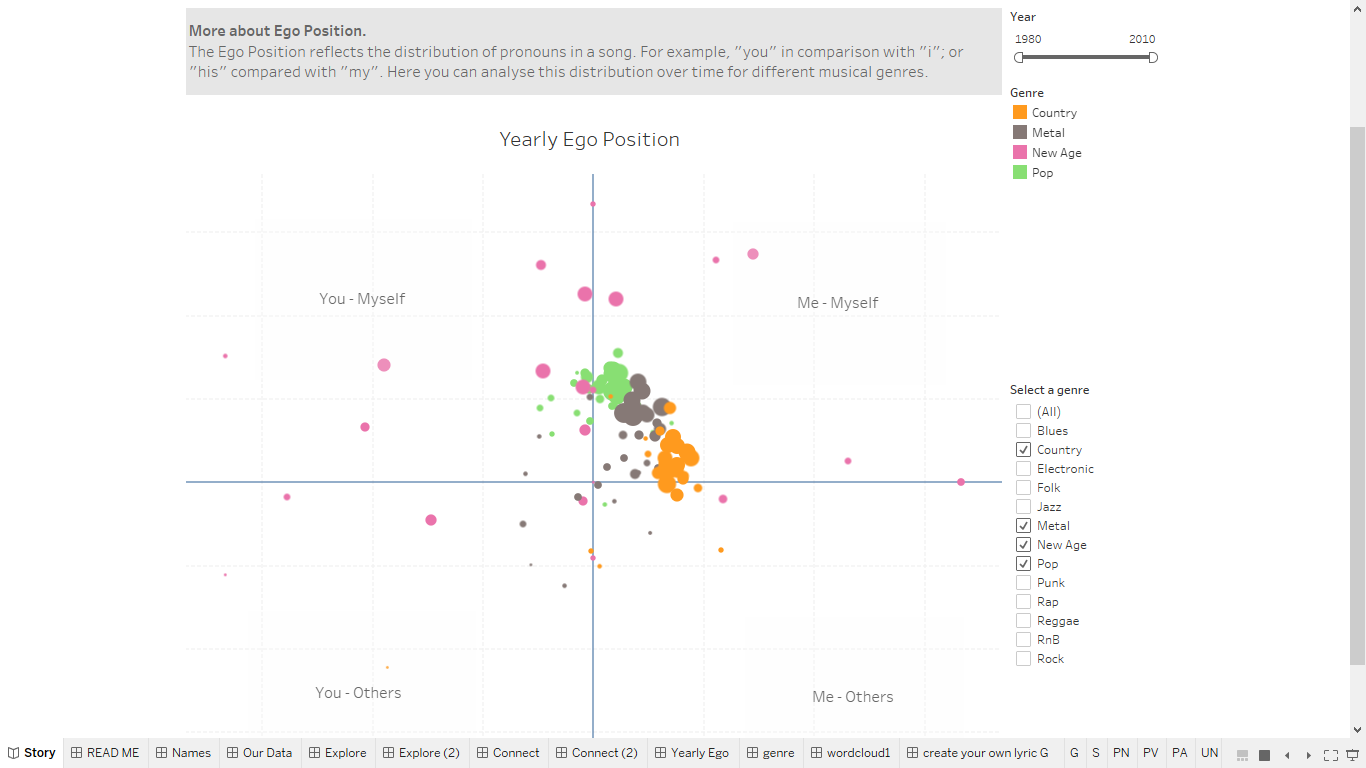
Alt 7 - Yearly Ego – Bars

Another alternative for the yearly ego chart is to simply use bar charts. This is a straightforward way for presenting the data, but as in the case of the time series, this leads to a complicated visualization.



Bars and time series depend strongly on the labelling of their axis, when displaying more than one category of data. Since our aim is to compare the behaviour of the music genres these alternatives do not fulfill our objective of providing insight at a glance.

The scatter plot that we use in our story allows for a clean presentation of trivariate data. We can encode the use of pronouns as the position on plane, and this opens the possibility of encoding more variables: years are encoded in circles, whose size encode the number of tracks for that year. The use of a plane also leads to a nice result: this reveals the clustering or dispersion of a genre. The more detailed information that the bars can provide is encoded as a tooltip in our visualization.



Alt 8 - Holoviews for Chords

Design Decisions

One alternative I explored thoroughly was to use the Holoviews framework for creating a similar chord diagram. The encodings would be similar. Nodes would represent word frequencies, edges would represent the bigram occurrences of each pair, and the number of edges would represent

The major difference would be in functionality discussed in detail below. Holoviews is less dependent on cursor movement and placement.

In summary, Holoviews is a far superior solution for mobile and for balanced datasets.

One alternative which I considered was to remove stopwords entirely. It would mostly solve the problem outsized nodes. However, many stopwords are very useful in creating grammatically correct titles. In a user’s travel around the diagram, they are simply too useful as origins and intermediate stops that I chose to leave them.

A picture containing ride, fireworks, device

Description automatically generated

Comparison

|  |  |  |
| --- | --- | --- |
|  | **Holoviews** | **Chord** |
| **Zoom** | Built-in support for mouse-wheel or pinch-to-zoom | Browser zoom only |
| **Pan** | Drag to pan | None |
| **Dynamic Scaling** | Yes | Yes |
| **Tap / Hover** | Clickable nodes highlight and lock the selection. Hover over tooltip. | Node highlighting on hover only. Hover over tooltip. |
| **Reset** | One click to reset to default state | Automatically reset when cursor moved off-screen |

1. In our setting, the tf-idf (term frequency – inverse document frequency) reflects the importance a word (term) in a song (document) in our whole collection of songs, i.e. our dataset. [↑](#footnote-ref-1)
2. The in-genre idf reflects the inverse frequency of a word in a music genre. [↑](#footnote-ref-2)