# **Comic Book Readership Archive (CoBRA)**

## **Exploring Comic Book Fan Mail across Gender and Geography**

[Michael Barron](https://iu.instructure.com/groups/254198/users/6000054), [Brett Steele](https://iu.instructure.com/groups/254198/users/5993917), [Tim Thompson](https://iu.instructure.com/groups/254198/users/5592060), [Yifan Xu](https://iu.instructure.com/groups/254198/users/3628875), [Ray Zupancic](https://iu.instructure.com/groups/254198/users/5907087)

### Project scope and significance

The project will produce a thematic visualization of published fan mail (taken from the time slice provided, 1961-1973) and consisting of several subcomponents that address the relevant themes in the dataset:

* Geographic origin
* Gender representation
* Comic book series popularity

These subcomponent visualizations will provide a unified view of comic book readership in the form of a world geospatial representation with a count analysis of letters published and a network analysis of letter writers and comic book series. The discrete quantifiable measures will include letter-writer breakdowns from U.S. states and countries and correlations among gender, location, and comic book series.

The significance of the CoBRA project lies in its examination of an area of popular cultural studies--comic book fan letters--that has not yet been adequately research or documented. A historical database of fan letters provides a unique resource for answering research questions about the feedback loop between production and consumption in the comic book genre.

### Sketches/preliminary visualizations

#### **Choropleth map of letters by U.S. state:**

#### **Sample network visualization of comic book series and letter writers:**

### Related work

The textbook Visual Insights1 contains exemplary visualizations and thematic presentations that are relevant to the CoBRA dataset, including the case studies "One Hundred Years of ISIS" (p. 262) and Visualizing NYC Hive (p. 268). The rough outline of our proposed CoBRA visualization is based on observing these thematic approaches to visualization and data encapsulation.

A search on some of the reference suggestions in the CoBRA project description reveals a few scholarly articles and materials relevant to the history of comic book readership and relevant to the formation of the CoBRA project. These references are useful for understanding the context around the project, but not in terms of auxiliary raw data or exemplary visualizations.

A search on comic-book readership online produces some more specific and contextually relevant data and visualization examples, including breakdowns of demographic-based and gender-based fan discussion and readership statistics. The Internet and Internet discussion and fan sites appear to be the modern replacement of older fan publications, and is naturally more expedient and simpler to harvest data from contemporary online sites and archives than from print sources. As an example, the Graphic Policy website details such statistics, as harvested from Facebook.2 depicts pie-charts and line graphs depicting comic book demographics).

Although interesting, it is not clear that online sites with present-day statistics provide any significant guidance to the task of characterizing, analyzing, and visualizing the historical data segment presented in the CoBRA project. An expanded scope that includes comparison to present-day statistics might be useful.

A previous version of the CoBRA dataset has been visualized by Anirudh R. A., Ashish Shendure, Brittany Yoder, Jae Eun Kum, and Shambhavi Dhargalkar.3 Our work will differ from theirs by focusing more on geospatial and network visualization techniques. Another exemplary visualization cited by Anirudh R. A. et al., which includes digital images and transcripts of comic book fan letters, can be found on the hosted TimeMapper website.4

### Data cleaning and preprocessing

#### **Gender data**

The approach to gender identification of names was based on using the Social Security Tables from 1880-1914 to identify registration trends. The tables were globbed into a master table and sorted by occurrence based on gender. The dominant count was used as the basis of projection: for example, in the case of the name "Billy," the ambigendrous ratio is given by: BillyMale381527 | BillyFemale5346.

Based on historic registrations, "Billy" was projected to be female in only 1.38% of occurrences. Therefore the code was designed to project the name as male (barring any obvious contraindications such as a person\_titlevalue of "Mrs."

#### **Geocoding data**

Geocoding data from the Google Map web services has been harvested and will be utilized in the project's next phase of visualizations. Using street-level geocoding from Google allows for more granular mapping and visualization versus centroid-based geocoding found in applications such as Tableau (also, Tableau allows for custom geocoding import, so one can gain very granular geocoding information and still use Tableau visualizations). The particular tools used included the geopy module and Python (registration with Google was necessary). A further advantage of Google is the byproduct of corrected, standardized addressing (i.e., address cleaning) for possible future incorporation into the dataset.

In some cases preliminary cleaning of address data was necessary (e.g., where the city and state were recognizable but the country was left null). This process preceded more granular address cleanup.

### Summary statistics

A summary of the dataset is presented in the table below. An initial count of names in the dataset yielded a total of 3,168 unique names. Person data is in the process of being further cleaned and reconciled, so this figure may change.

| Comic book series | Comic book issues | Fan letters | Countries | Unique names of letter writers | Gender of letter writers (inferred) |
| --- | --- | --- | --- | --- | --- |
| 11 | 610 | 3703 | 38 | 3237 | | Male | Female | Null | | --- | --- | --- | | 2863 | 266 | 108 | |

### Visualizations

Our initial approach to visualizing the CoBRA data has focused on the variables of gender and geographic location (currently limited to US states). Methodologically, we have explored three techniques: statistical visualization, geospatial visualization, and network visualization.

#### Statistical

To analyze the fan letter data, a Pivot Table was created in Excel of all of the CoBRA data. For rows, the series\_name and issue\_id data fields were selected. For the value field, the count of issue\_id was selected. This provided a count of fan letters for each individual issue and also provided an overall count of fan letters for each series.

To create a frequency distribution of fan letters and issues, columns were numbered from the lowest value of fan letters per issue (1) to the highest value (71). A binary =if() statement matrix was created, referencing the values in the pivot tables to those of the numbered columns. If the number of fan letters was equal to the column number, the cell would show 1. If it did not, the cell would show 0. The columns were then totaled to obtain the number of individual issues receiving each number of fan letters. Once these values were obtained, a scatterplot was created with the x-axis representing the number of fan letters and the y-axis representing the number of issues.

Referencing the same pivot table, a histogram of the total number of fan letters for each comic book series was created, as well as one for the average number of fan letters per issue in each series.

An additional histogram was created to represent the breakdown of gender by state.

#### **Links to preliminary statistical visualizations**

##### [Frequency distribution of fan letters and issues](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/issue-freq.png)

##### [Total fan letters by series](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/letters-series.png)

##### [Average number of fan letters per issue, grouped by series](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/letters-issue.png)

##### [Bar graph of gender totals by state](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/state-gender.png)

#### Geospatial**: choropleth map of the proporation of female-to-male letter writers by US state**

To visualize the female-to-male proportion of CoBRA letter writers, R was utilized to analyze, clean, and reshape data.5 From there, shape files from ArcGIS were used to map the female-to-male proportion of US letter writers. The data was read from R and persons were filtered by id\_person\_dim to identify unique authors and then grouped by gender and state to aggregate gender totals.

The female-to-male proportion was calculated from these totals and map tools were used to map, fortify, and merge state dimensions to state names from the data file. This data was plotted as a Choropleth Map with ggplot2. The higher the proportion of female-to-male values, the darker the shade of red.

#### **Link to preliminary geospatial visualization**

##### [Choropleth map showing the proportion of female-to-male letter writers by state](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/f2m.png)

#### Network**: force-directed graph of comic book series and letter writers, color-coded by gender**

To visualize the connections among letter writers and comic book series, a force-directed network graph was encoded as a GraphML/XML file and rendered in Gephi. The BaseX XML database and XQuery interpreter was used to convert the CoBRA CSV data file into an internal XML representation. An XQuery script was then written to produce node and edge lists using GraphML notation. Node values were color-coded to distinguish between series nodes and letter-writer nodes, the latter of which were color-coded by gender (male, female, null).

The resulting GraphML file contained a total of 3248 nodes and 3482 edges. Once imported into Gephi, a series of network statistics was calculated, including degree, betweenness centrality, and modularity class. These statistics are still being explored and have not yet been incorporated into additional visualizations. The network was laid out using Gephi's implementation of the Fruchterman Reingold force-directed layout, with the following parameters:

| Area | Gravity | Speed |
| --- | --- | --- |
| 10000.0 | 7.0 | 10.0 |

Nodes were then sized by degree, on a scale of 8 to 50, and proportionally sized node labels were added. Labels were then filtered to show only those labels whose nodes had a degree greater than 9.

#### **Link to preliminary network visualization**

##### [Force-directed network graph of comic book series with nodes color-coded by gender](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/gephi.xml)

### Insights

The preliminary phase of visualization and analysis has revealed some basic insights regarding the contours of the data:

* *Fantastic Four* (901 issues) and *The Amazing Spider-Man* (1092 issues) represented the greatest portion of the data by series, but *The Avengers* and *The X-Men* had the highest average of number of letters per issue (8.28 and 7.69 letters per issue, respectively).
* All 50 fifty US states and several other countries were represented in the data; California and New York were by far the biggest contributors.
* The letter writers in the dataset were overwhelmingly male, but the most active letter writer (Shirley Gorman, with 17 letters) was female.
* Alaska, New Hampshire, Idaho, Arizona, and New Mexico were the only states with a proportion of female-to-male letter writers greater than 25% (1:4).
* By region, the Pacific Northwest had the highest proportion of female-to-male letter writers.

The next phase of visualization and analysis should yield additional insights into the connections among the people, places, and comic books represented in the dataset.

### Validation and redesign

#### **Deduplication**

During the data clean-up process, many records were observed to have duplicate names but different id\_person\_dim identifiers and different location information. Given the selective nature of the data, it is still very possible that these are the same individuals writing from different locations. For example:

| id\_person\_dim | Name | Address |
| --- | --- | --- |
| 2902 | Adams Richard | 102 Bryan Ct. Apt. 101 Laurel MD US 10810 |
| 2458 | Adams Richard | Rt. #1, Box 439 Oakiland MD US 21550 |

In this case, the issue dates are two years apart. It's reasonable to assume that Adam could have moved from Laurel, Maryland, to Oakland, Maryland. The following disambiguation algorithm was decided upon: if a name is the same, the address is within the same state, and the difference between publication dates is within 2 years, then identifiers for this person will be merged.

In terms of workflow, all duplicate names were extracted with their identifiers to a new file (901 records) and merged according to the disambiguation algorithm. Finally, the cleaned file will be used to generate more visualizations whose data points are less dispersed.

Conversely, there are also data validation issues for cases in which names are identical or similar, but the location is the same. For example:

| id\_person\_dim | Name | Address |
| --- | --- | --- |
| 2472 | Burton Mike | 1149 Greenwood St Bellefontaine OH |
| 2473 | Silvers Mike | 1149 Greenwood St Bellefontaine OH |
| 2583 | Burton Mike | 1149 Greenwood St Bellefontaine OH |

#### **Geocoding**

Centroid-based geocoding was augmented with street-level geocoding from Google’s geocoding web services, in order to potentially produce more informative geospatial visualizations. Several issues were encountered:

1. street addresses not always clean enough to pass through Google
2. street addresses that use non-standard designations (e.g., "barracks" rather than street)
3. street designations that Google silently passes to a centroid based on city, state, country

Addresses in these categories had to be flagged and manually inspected or modified. More generally, it is not clear that street-level geocoding adds much to the visualization (does a dot on South Denver inform more than a dot on centroid Denver on a small map?). At the same time, there is value in clean addresses in the overall dataset (the project curator might want a cleaned address scheme going forward), but this is not one of the deliverables asked for in the project scope.

#### **Gender data**

The strategy for determining gender from author names involved starting with Social Security Administration (SSA) Tables from 1880 to 1914, and using these to identify author gender based on historic registration trends listed in the SSA data.

In the case of ambigendrous names, a programmatic mitigation is possible. The dominant gender is deteremined based on preponderance of registrations per the time interval of interest. For example, "Billy" was projected to be female in only 1.38% of occurrences. Therefore the code was designed to assign the name as male--taking into account any obvious contraindications such as a title of "Mrs."

#### **Uniqueness criteria**

One aspect of analysis that came up in the midst of initial visualizations was the question of relating unique versus aggregated data points for a single letter writer. Many questions that one may attempt to visualize have an implicit assumption of uniqueness or non-uniqueness that must be addressed as part of the analysis. For example, in a geospatial visualization, if the same writer produces a significant number of letters from a single location, should that be part of the visualization or should that writer count as weight 1. As an extreme consider, for example, if a single writer from a town in Australia were to write 100 letters: the visualization might lead one to believe that a certain comic book had a significant following in Australia, which would be misleading. Uniqueness and distinctness are important to understand and relate.

### Challenges and opportunities

#### **Visualization design and data modeling**

One of the unique challenges of the CoBRA dataset lies in finding a visualization approach that does justice to the number of connections and relationships within the data. The data was created using a relational database schema, so many of those linkages are present, but they may not be easy to highlight and visualize.

The issue of connectedness seems especially important to the CoBRA data. One of the insights that has emerged during the course of analysis is the importance of comic books as a locus of community and connection across time and space. There are several instances of multiple letters from different individuals located in the same place. Take the case, for example, of Ann, Valerie, Wendy, Barbara, and Mary, who all wrote letters from 1 Sundial Lane in Birmingham, England, to the Fantastic Four fan page in April and May of 1968. There are other cases of letters that appear to come from the same household, apartment building, or military barracks, suggesting that comics can form an "imagined community" of readers (to borrow Benedict Anderson's phrase).5

To address these challenges and attempt to convey a sense of the complex interplay of time, space, and relationships in the dataset, a different approach to data modeling was adopted. A reduced version of the master data table was created by removing columns that contained administrative data. This core data table was uploaded to the Karma data integration tool,6 a platform for converting relational data into linked data using Semantic Web principles, vocabularies, and protocols. In Karma, columns and rows were mapped to linked data vocabularies such as Sean Petiya's Comic Book Ontology.7

The RDF linked data generated by Karma was then further cleaned and converted to RDFa (RDF in Attributes) for embedding into HTML Web pages.8 A primary component of the CoBRA Group 2 visualization deliverable will focus on the ability to explore and visualize data linkages through online exploration and navigation. The preliminary architecture of the project website reflects an attempt to implement "5-star linked data," following the principles outlined by Tim Berners-Lee.9

An example of what this approach might look like in practice is provided in [this page for Shirley A. Gorman](http://ivmooc-cobra2.github.io/home/tat2/datasci/ivmooc/client/ivmooc-cobra2.github.io/assignments/4/fans/530/), the most profilic letter-writer in the dataset.

### Citations

1Börner, K., and D. E. Polley. 2014. *Visual Insights*. Cambridge, MA: MIT Press.

2http://graphicpolicy.com/2015/09/01/demo-graphics-comic-fandom-on-facebook-14/

3http://ashishshendure.com/projects/ivmooc\_cobra\_2015/

4http://timemapper.okfnlabs.org/anon/ik4x3a-timemapper-comics-readership-example

5The following R packages were used: RCurl, dplyr, reshape2, ggplot2, map tools, scales, and RColorBrewer.

5Anderson, B. 1983. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London, England: Verso.

6http://usc-isi-i2.github.io/karma/

7Petiya, S. 2015. "Comic Book Ontology." https://comicmeta.org/cbo/

8https://www.w3.org/TR/rdfa-primer/

9http://5stardata.info/en/

Related work

The textbook Visual Insights1 contains exemplary visualizations and thematic presentations that are relevant to the CoBRA dataset, including the case studies "One Hundred Years of ISIS" (p. 262) and Visualizing NYC Hive (p. 268).

A search on comic-book readership online produces some more specific and contextually relevant data and visualization examples, including breakdowns of demographic-based and gender-based fan discussion and readership statistics. The Internet and Internet discussion and fan sites appear to be the modern replacement of older fan publications, and is naturally more expedient and simpler to harvest data from contemporary online sites and archives than from print sources. As an example, the Graphic Policy website details such statistics, as harvested from Facebook.2 depicts pie-charts and line graphs depicting comic book demographics).

A previous version of the CoBRA dataset has been visualized by Anirudh R. A., Ashish Shendure, Brittany Yoder, Jae Eun Kum, and Shambhavi Dhargalkar.3 Our work will differ from theirs by focusing more on geospatial and network visualization techniques. Another exemplary visualization cited by Anirudh R. A. et al., which includes digital images and transcripts of comic book fan letters, can be found on the hosted TimeMapper website.4