University of British Columbia
CPSC 547 - Information Visualization
Tamara Munzner

# Visualizing Students Migration in Elementary and Secondary Schools in São Paulo/Brazil

Proposal

Carolina Roman Amigo
carolamigo@gmail.com

Wenqiang Dong
wdong@cs.ubc.ca

9th November 2015

# 1. Domain, Task and Dataset

## 1.1 Domain

This project domain is elementary and secondary education in the state of São Paulo, Brazil. As in Canada, elementary and secondary education in Brazil consists of twelve years of education for children aged from 6 to 18 years (grades 1 to 12). There are both public and private schools available, the private ones outperforming public ones regarding education quality [18, 5]. This is well known by parents and affects their school choices; according to Estevan (2015), it is possible to establish a link among public education quality in Brazil and number of enrollments in private schools. When quality of public education increases, there is a decrease in the number of enrollments in private schools.

In the state of São Paulo, 35.5% of the schools were private in 2014. These schools find a large number of families willing to pay for their children's education, as the state has the highest income per capita of the country. As any business, private schools share this market and compete with each other for students to survive. Factors such as home-school distance, increase of tuition fees, and periods of economic growth or decline may influence the number of enrollments. Another high impact factor is the number of students from a given school who were able to get to the top three universities of the state, considering that the only admission criteria is a high score in a standardized test. Schools that specialize in training students for that test are most likely to have a large number of new students enrollments at high school, since these are the years that precede college.

As outlined above, students may migrate from school to school for several reasons. Understanding these migration patterns is useful for both government and private schools, because it may help them identify issues and potential areas of improvement. For example, a steady increase of student's migration from public to private schools may be a warning to the government that the quality of public education is decreasing. For a private school, if they are consistently losing students around the 9th grade for another private school, this may be a warning that they are not investing enough effort in preparing students for the college standardized admission test.

## 1.2 Tasks

We have two different types of stakeholders interested in this dataset (schools and government). Both are interested in understanding migration of students, but in different granularity levels; while for the government is interesting to get an overview of migration among schools, schools are mainly interested to understand the specific migration flow from and to it. They will use the visualizations to compare migration

among grades and schools, spot outliers and understand the migration network. Examples of the specific tasks that will be supported by our proposed visualization are:

- Task for schools
  - Is there any particular grade in which migration is more intense (losing/gaining more students)?
  - To which schools are their students going to/coming from?
  - Is it possible to identify a pattern in the geographic location of the main competitor schools?

- Task for government
  - Are students migrating from public to private schools (or vice versa)?
  - Which schools are gaining/losing more students?
  - Which grade is gaining/losing more students across schools?

## 1.3 Dataset

We are using data from the Educational Census, which is released every year and it is public available at the government website. For each student of the country, this census shows in which school they were studying in that year, in which grade they were, and the school type (private or public), among other related information. It is a huge dataset; just for the state of São Paulo we have 10, 581, 500 students and 28, 718 schools, as of 2014. The actual number of students we are going to work with is larger as we are going to use data from census since 2012 until 2014, in the order of 10.000.000 rows each.

We are using two tables from the census dataset: “Enrollments” table which has enrollment_id as a primary key, and has all students enrollments for that census year; and “Schools” table, which has school_id as a primary key and lists all the schools in the state, with names, latitude and longitude information, among others. Herein is a table summarizing the fields we are going to use in our project.

Table 1 – Dataset fields selected from the source files for building the visualizations

| **Source table** | **Field name** | **Description** |
|---|---|---|
| school | school_id | primary key, id for each school. |
| | school_name | the name of the school |
| | school_city | foreign key, city code of the school |

| | | |
|---|---|---|
| | school_district | district code of the school |
| | post_code | postal code of the school |
| | latitude | geographical coordinates of the school |
| | longitude | |
| | school_status | status of school (active, inactive) |
| enrollment | year | enrollment year of a student, from 2012 to 2014 |
| | enrollment_id | primary key, id for each enrollment item |
| | student_id | id for each student |
| | education_grade | student's educational grade |
| | school_id | id for each school |
| | school_city | city code of the school |
| | school_type | type of the school, public (federal, state, city) or private |

# 2. Personal Expertise

Carolina used a small subset of this dataset to build one visualization for the second part of a visual analytics course she took at the Vancouver Institute for Visual Analytics (VIVA). The second part of the course consisted in a project within a data visualization company, in order for the students to get in contact with real world visual analytics challenges. The company chosen was based in Brazil and was developing a product for georeferenced marketing for private schools. She focused on migration to and from one grade of a specific school in the state of São Paulo, in one year. As she has a background in graphic design, her main interest was to explore visual encoding options using a mockup tool (Adobe Illustrator CC). She performed data gathering and cleaning as well as light coding to mockup purposes using Google API. Her inputs were accepted as suggestions for the final product, which is still in development.

Dylan never saw the dataset before and this is his first time doing a geographical visualization. He doesn't know the area at all, but this seems like an interesting problem. He has experience with graph drawing. His research focuses on how to draw small scale graphs satisfying multiple aesthetic criteria and how to increase the speed of drawing large scale graphs by heuristic initial placement of the nodes.

# 3. Proposed Infovis Solution

We use the three-part analysis framework for a vis instance to explain our infovis solution [12]. Figure 1 summarizes what data the user will see in our visualizations. We are proposing a dashboard (Figure 6) which will combine the following dataset types:

- Network, showing the migration flow to and from a selected school.
- Fields, bar and column charts showing the total number of students leaving and entering each school per grade in a given year (overview panel) and students migrating to and from a given school, as well the flow balance (school view panel).
- Geometry, showing geographic location of schools in a map (school view panel).

The dataset we have access to is static. We will show both categorical attributes (grades) and quantitative attributes (total number of students migrating to and from schools, balance of migration flow). The ordering direction will be both sequential (crescent list by migration total) and diverging (balance of migration flow).

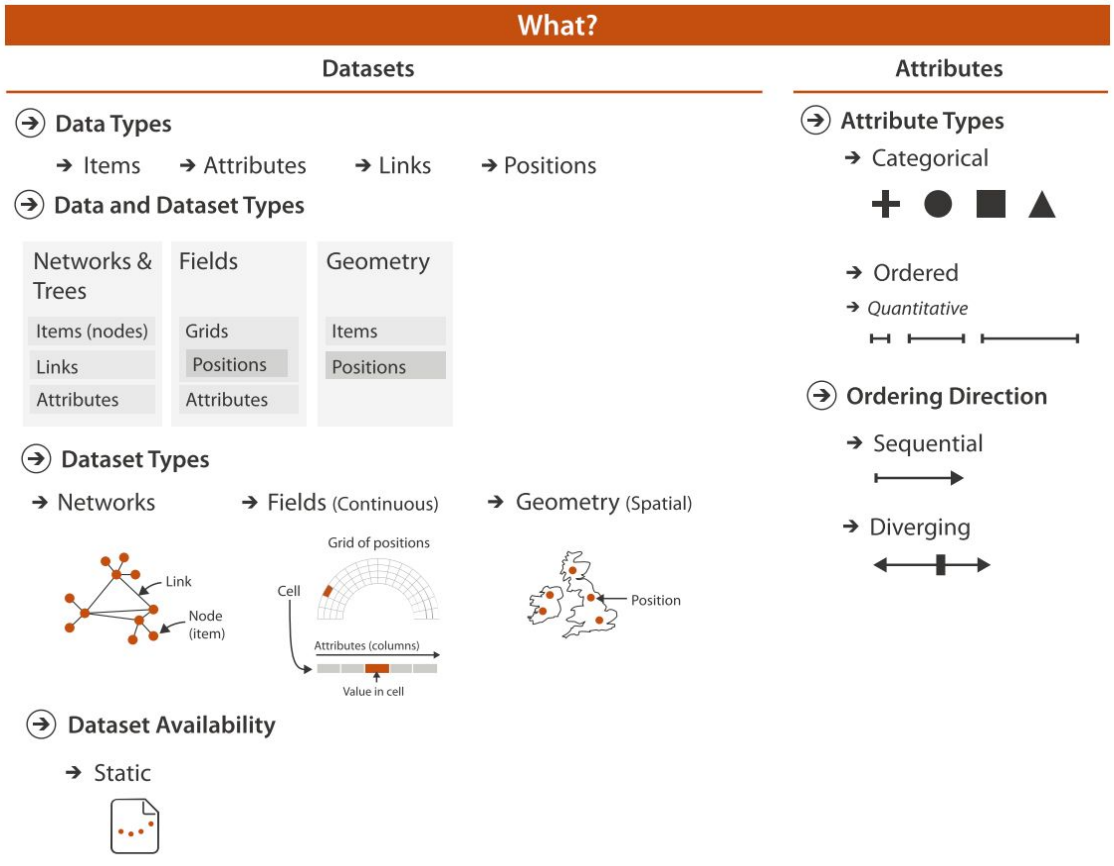


Figure 1 - What data the user will see in our visualization (adapted from Munzner, 2015).

Users will be probably using our visualizations to compare migration among grades and schools, spot outliers and understand the migration network (Figure 2). As we are dealing with a large dataset, users will also use derived dataset to get a summarized view of the migration flow, for example total number of students schools lost in a given year. Although we don't have an extensive sequence of census years to analyse in order to allow users to discover migration trends over time (are are using only three years), users can discover trends among schools, for example, a specific grade which always shows a large outflow of students.

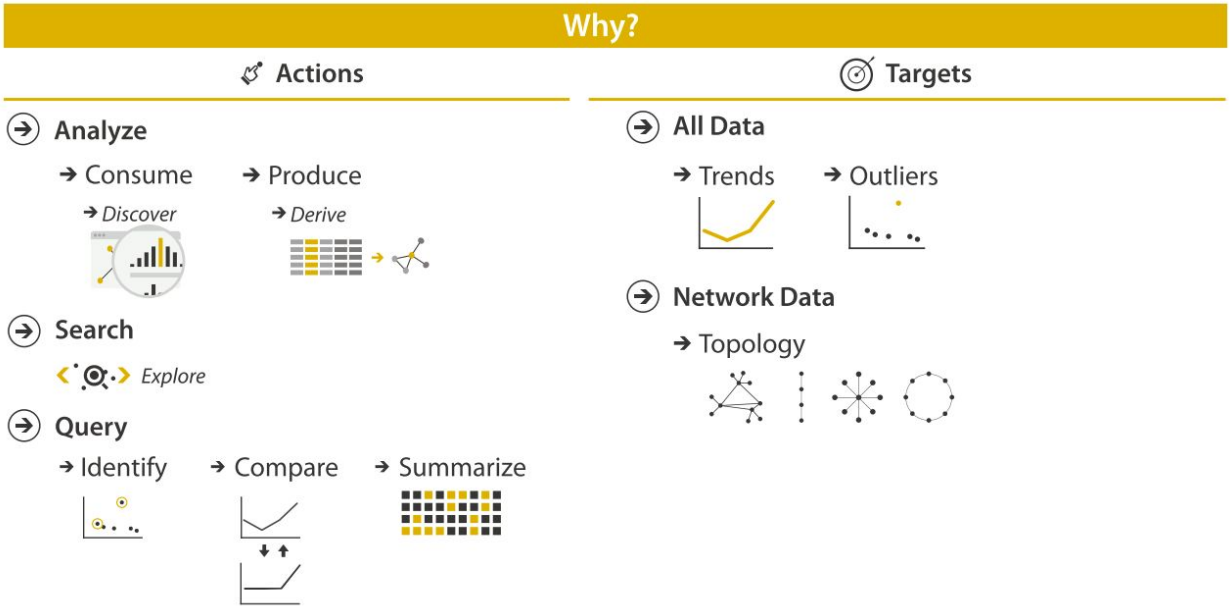


Figure 2 - Why the user will intend to use our visualization (adapted from Munzner, 2015).

Figure 3 summarizes how the visual encoding will be constructed. To cope with the size of the dataset, we opted to make use of a combination of navigation, facets and filtering. Beforing even visualizing the "Overview" panel (Figure 4), users have to define a pre filtering informing year, focus school type, flow type (inflow, outflow, balance), and type of competitor schools (private, public, both). Once in the "Overview" panel, users can scroll it in order to see the whole list of schools in the bar chart. Users can filter the list by number of migrating students, and also by grade. They can order the overview list in crescent or decrescent order by the total or by grades per school. Color hue is used to differentiate grades. Once users find a school of interest in the "Overview" panel, they can select that school in order to see its detailed dashboard ("School View" panel).

In the "School View" panel, we will have a dashboard offering three facets of the dataset: a parallel coordinates graphic showing the size of flow to and from a school, per grade; a column chart showing the total inflow, outflow or flow balance per grade; and a map showing geographic migration, also per type of flow. Users will be able to select a given grade to highlight it, which will be achieved using color saturation. Flow

sizes will be encoded using stroke width in the parallel coordinates graphic, and school types will be encoded with different marks shapes in the map.

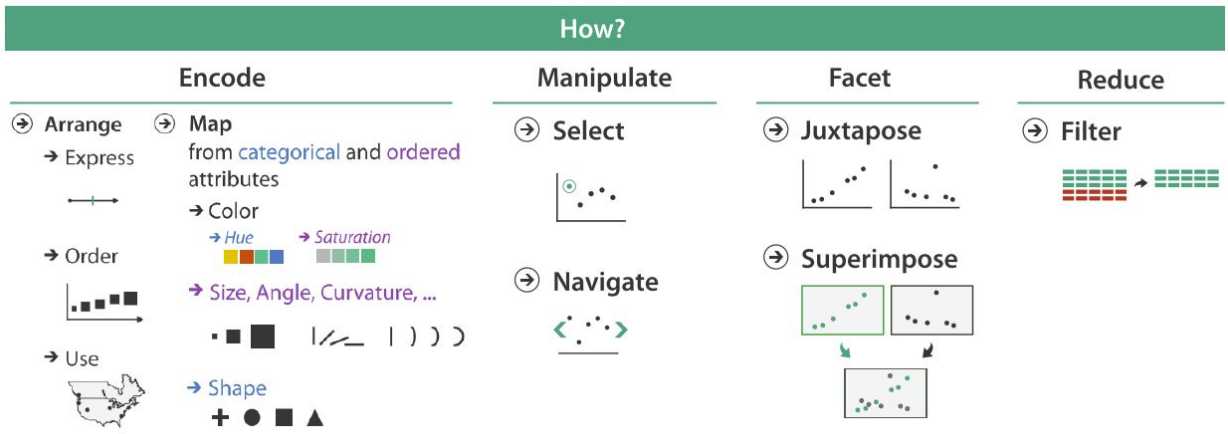


Figure 3 - How the visual encoding will be constructed in terms of design choices (adapted from Munzner, 2015).

# 4. Scenario of Use

## 4.1 Private school trying to understand students losses

A school director is noticing the number of enrollments falling in recent years. He would like to know the reason for students to leave his school, and it would help to know where their students are going to. So he decides to use the migration vis tool we are proposing. As he is interested in a specific school, he clicks on the “School View” panel (Figure 6), and searches for his school using the search field. After that, he selects the year he is interested in (2013), and sets to “private” the competitor type of school his school is receiving/sending students to, because this is the type of school he shares the market with. The visualizations appear in the dashboard a few seconds after he selects the competitor type.

On the parallel coordinate view, he has an overview of the inflow/outflow of students per grade. In the middle axis, there are the grades; on the left one, the schools from which his school is receiving students; on the right axis, the schools for which his school is sending students to. Hovering the mouse over a grade, he can highlight it. He notices that a large amount of students are leaving on the 9th grade, mainly to three other schools. So he clicks on 9th grade to select it, and the the information about that grade is highlighted on the two other views too. On the column bar chart at the top right, he selects “balance” and confirms that he is losing a large amount of students in the 9th grade, more than in any other grade.

On the geographic flow view at the bottom right, he can see the geographic location of the schools for which he is sending students in the 9th grade. As they are spread all over the city, geographic location doesn't seem to be the reason for its students to prefer those schools. So he takes notes of the names of the three competitors and searches for them on the internet. He notices that all of them are offering a special high school curriculum to prepare students for the admission test for universities, something his school doesn't do, and this is a strong hypothesis for his students to be leaving at this specific grade. With that information in hands, he can now discuss a plan of action with the school board.

## 4.2 Education department trying to understand students evasion for private schools.

An analyst from the Education Department of the State of São Paulo is working on a project to improve education quality in elementary and secondary public schools. He wants to identify specific schools in need of interventions. He knows that middle-class parents with limited budget for paying for their children's education use a combination of public and private education. So he would like to know which public schools are having the largest evasion to private ones, and if there is a pattern of evasion, for example one specific grade at which the evasion is higher.

He decides to use the migration vis tool we are proposing. As he wants to have an overview of schools, he selects the "Overview" panel (Figure 4). The he selects the year he is interested in (2013), the type of focus schools he is exploring (public ones in this case), and the flow type (outflow as he is interested in evasion). Finally he selects the type of school students are going to (private). A bar chart with all grades evasion for each public school appears in the screen after a few seconds. He can scroll the list and also filter it, as it is large. He chooses to filter the total migration per school to more than 25 using the slider at the filter bar, as he is only interested in the largest migrations. Then, he clicks in the arrows at the total column header in order to order the list of schools in decrescent order.

He can see now which schools are losing more students for private schools in the state. He can also identify a pattern: losses seem to be higher around 1st grade and 9th grade. By clicking the "x" in grades headers, he can hide the columns he is not interested in, in order to focus on the grades with greater losses. He can order the list by grades also, clicking the arrows in the grades columns headers. As the first school of the list seems to have the largest evasion of students for every grade, he clicks on it to see details on the "School View" panel (Figure 6, panel already explained in scenario 4.1).

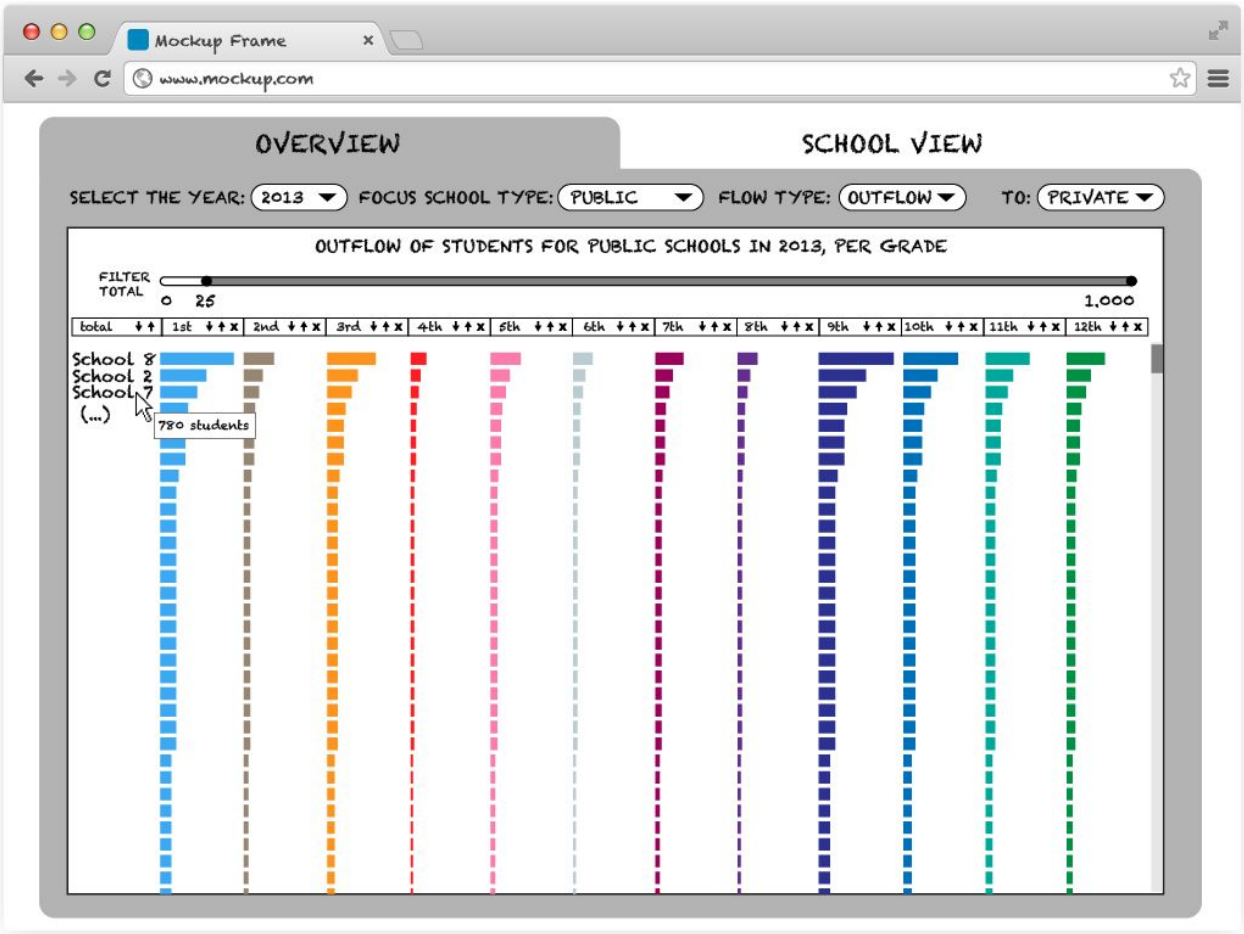


Figure 4 - “Overview” panel mockup, offering an overview of outflow and inflow of students to schools, per grade.


Figure 5 - Details of dropdown menus on the overview panel.

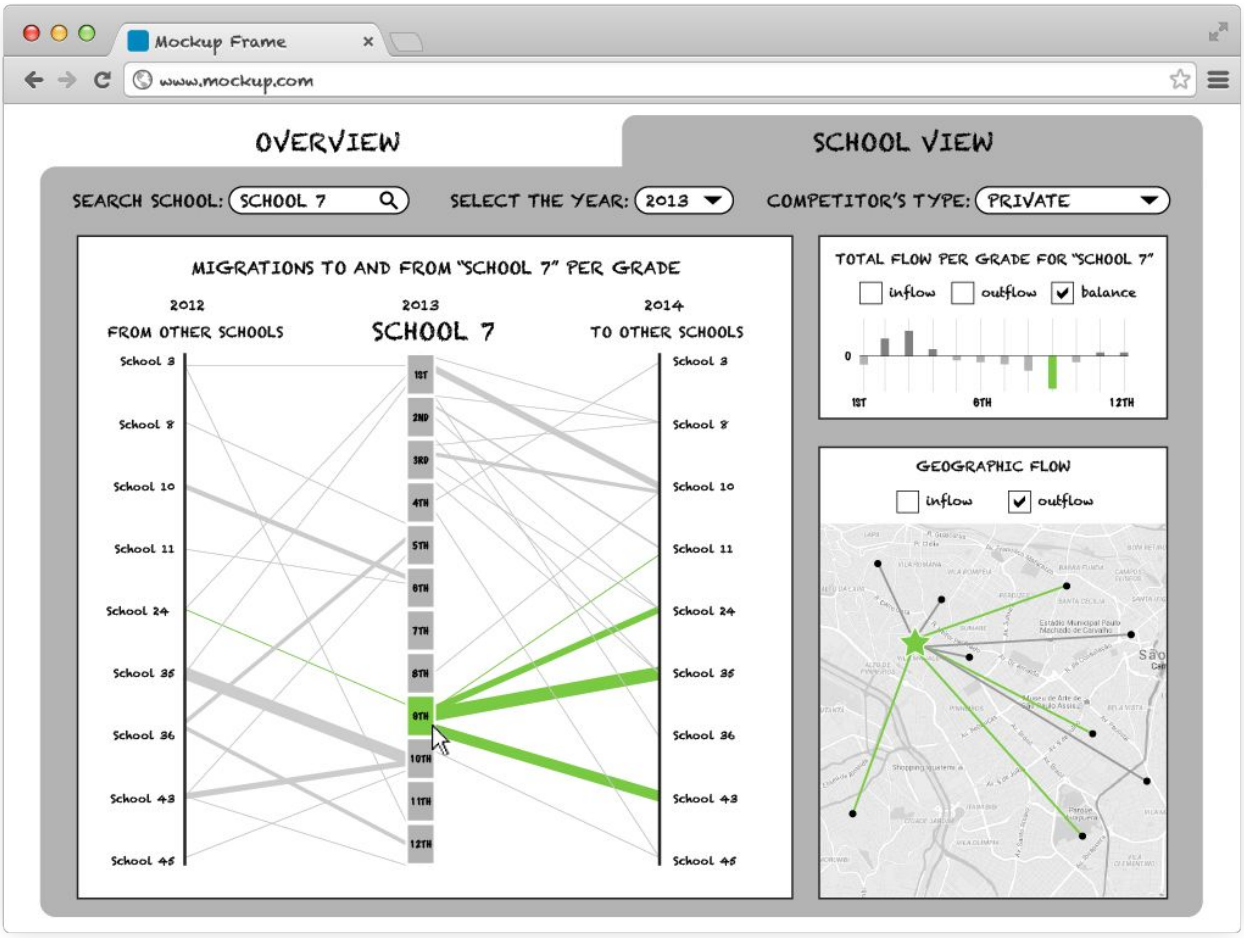


Figure 6 - "School view" panel mockup, containing a migration dashboard for a selected school.

# 5. Proposed Implementation Approach

We will design a web-based visualization system. For the web server software, we will use Apache HTTP Server. The data will be stored in a MySQL database. The programming languages that will be probably useful for this project are JavaScript, HTML, PHP and MySQL. In addition, we will use Python for data pre-processing. The visualization system will run on the Mac OS platform. We might use D3.js to help in visualization design and interaction, and Google Maps APIs to customize the map and plot schools locations on it.

# 6. Milestones and Schedule

| Date | Milestone |
|---|---|
| 11/9 | Finish the proposal |
| 11/13 | Finish the front end design for the Overview interface |
| 11/18 | Finish the back end coding for the Overview interface<br>Test Overview interface |
| 11/22 | Finish the front end design for the School View interface |
| 11/23 | Finish the Status Updates report |
| 11/26 | Finish the back end coding for the parallel coordinates plot |
| 12/3 | Finish the back end coding for the flow map |
| 12/7 | Finish the back end coding for the bar chart |
| 12/11 | Finish linkage among the above three windows<br>Test School View interface |
| 12/14 | Finish and test the visualization system development<br>Prepare slides for the presentation |
| 12/15 | Finish final presentation |
| 12/18 | Finish final report |

# 7. Previous Work

Origin-destination datasets have been getting more accessible in recent years, e.g. flows of people, animals, traffic, knowledge, disease, etc. Typically, the structure of this kind of data is complicated and its scale very large. To help people get a direct understanding of the origin-destination data effectively, visualizing data flows have been drawn more attention.

The first example of geographic flow visualization was produced by Ravenstein [17]. He drew the movement of people around Great Britain and Ireland by means of a series of single headed arrows, crossing county boundaries and typically flowing

towards major urban centres. Then, 74 years later, the Chicago Area Transportation Study produced the first computer based flow mapping example [3]. Since then, many visualization methods have been proposed.

Boyanin et al. present Flowstrates to help users perform spatial visual queries and analyze the changes over time [1]. They display the origins and the destinations of flows in two separate maps and the changes over time of the flow magnitudes are represented in a separate heatmap view in the middle. Querying, filtering, ordering and grouping techniques are used to help interactive exploration.

In order to prevent losing details and introducing arbitrary artefacts in the visual representation, Wood et al. propose a method which maps the origin-destination vector as cells, in contrast to lines used by other methods [20]. They project geographic data on a set of spatially ordered small multiples by constructing a gridded two-level spatial treemap.

Rae uses flow density maps to visualize a large migration matrix from the UK's 2001 census [16]. Similarly, Gilbert et al. use statistical summaries of spatial association to visualize the movements of animals infected by bovine tuberculosis on the flow density maps [7]. They explore the association between bovine tuberculosis occurrence and the predictors by conducting a stepwise multiple logistic regression analysis of 2002 and 2003 bovine tuberculosis distribution data.

Verbeek et al. propose a method based on spiral trees, a type of Steiner tree which uses logarithmic spirals, to visualize flow maps [2]. They integrate edge-bundling to their algorithm and compute crossing-free, merge smoothly, and naturally cluster flows. The high-quality flows are produced by minimizing a global cost function which consists of obstacle cost, smoothing cost, angle restriction cost, balancing cost and straightening cost.

Phan et al. present a method to draw flow maps based on hierarchical clustering [15]. Their system consists of two phases: layout phase and rendering phase. They use distortion to ensure the nodes are well spaced but still preserve their relative positions to the neighbours. The edges are merged based on their destinations using hierarchical clustering. They use the spatial information given by the hierarchical clustering to do edge routing to avoid edge crossings.

Edge-bundling algorithms based on hierarchical information [9], geometry information [4], force-directed algorithm [10] and quadtree structure [11] are also used to visualize origin-destination data because they can reduce visual clutter by merging edges.

In addition to the above mentioned single-view methods, Guo uses multi-view displays to visualize migration flows [8]. The methodological framework consists of methods for hierarchical regionalization, flow mapping, multivariate clustering and visualization. The multi-view displays use a self-organizing map, parallel coordinate plot, and a flow map to present flow structure, multivariate information, and spatial patterns at the same time.

To explore the link between different objects, parallel coordinates technique may be used. It has been applied to many multidimensional problems and has been incorporated into many commercial and public-domain systems, such as WinViz [14] and XmdvTool [19].

Fua et al. enhance the parallel coordinates technique by developing a multi-resolutional view of the data via hierarchical clustering [6]. They make use of variable-width opacity bands to represent the information at a node. They also use a proximity-based coloring scheme to guarantee that data and clusters from similar parts of the hierarchical structure are shown in similar colors.

Novotny et al. integrate focus+context visualization in the parallel coordinates [13]. After binning the data into different levels of detail, they can visualize context information at several levels of abstraction while leaving enough visual resources for the outliers and for the data items in focus.

# 8. References

[1] Boyandin, Ilya et al. “Flowstrates: An Approach For Visual Exploration of Temporal Origin-Destination Data.” *Computer Graphics Forum* 30.3 (2011): 971–980. Web.

[2] Buchin, K., B. Speckmann, and K. Verbeek. “Flow Map Layout Via Spiral Trees.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011): 2536–2544. Web.

[3] *Chicago Area Transportation Study: Final Report*. Chicago: CATS, 1959. Print.

[4] Cui, Weiwei et al. “Geometry-Based Edge Clustering For Graph Visualization.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008): 1277–1284. Web.

[5] Estevan, Fernanda. “Public Education Expenditures and Private School Enrollment.” *Canadian Journal of Economics/Revue canadienne d'économique* (2015): Web.

[6] Fua, Ying-Huey, M.o. Ward, and E.a. Rundensteiner. “Hierarchical Parallel Coordinates for Exploration of Large Datasets.” *Proceedings Visualization '99 (Cat. No.99CB37067)* (1999): 43-50. Web.

[7] Gilbert, M. et al. “Cattle Movements and Bovine Tuberculosis in Great Britain.” *Nature* 435.7041 (2005): 491–496. Web.

[8] Guo, Diansheng. “Flow Mapping And Multivariate Visualization of Large Spatial Interaction Data.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009): 1041–1048. Web.

[9] Holten, D. “Hierarchical Edge Bundles: Visualization Of Adjacency Relations in Hierarchical Data.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 741–748. Web.

[10] Holten, Danny, and Jarke J. Van Wijk. “Force-Directed Edge Bundling For Graph Visualization.” *Computer Graphics Forum* 28.3 (2009): 983–990. Web.

[11] Luo, Sheng-Jie et al. “Ambiguity-Free Edge-Bundling For Interactive Graph Visualization.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 18.5 (2012): 810–821. Web.

[12] Munzner, Tamara. *Visualization Analysis and Design*. (2014). Print.

[13] Novotny, M., and H. Hauser. “Outlier-Preserving Focus Context Visualization In Parallel Coordinates.” *IEEE Trans. Visual. Comput. Graphics IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 893–900. Web.

[14] Ong, Hwee-Leng, and Hing-Yan Lee. “Software Report: Winviz—A Visual Data Analysis Tool.” *Computers & Graphics* 20.1 (1996): 83–84. Web.

[15] Phan, Doantam et al. “Flow Map Layout.” *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (2005): 29. Web.

[16] Rae, Alasdair. “From Spatial Interaction Data to Spatial Interaction Information? Geovisualisation and Spatial Structures of Migration from the 2001 UK Census.” *Computers, Environment and Urban Systems* 33.3 (2009): 161–178. Web.

[17] Ravenstein, E. G. “The Laws Of Migration.” *Journal of the Statistical Society of London* 48.2 (1885): 167. Web.

[18] Vandenberghe, V., and S. Robin. “Evaluating The Effectiveness of Private Education across Countries: a Comparison of Methods.” *Labour Economics* 11.4 (2004): 487–506. Web.

[19] Ward, M.o. “XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data.” *Proceedings Visualization '94* (1994): 326-333. Web.

[20] Wood, Jo, Jason Dykes, and Aidan Slingsby. “Visualisation Of Origins, Destinations and Flows with OD Maps.” *The Cartographic Journal Cartogr. J.* 47.2 (2010): 117–129. Web.