Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **2**, Number **2**, April **2017**
doi:10.3934/bdia.2017003
pp. **107–118**

RENDERING WEBSITE TRAFFIC DATA INTO INTERACTIVE TASTE GRAPH VISUALIZATIONS

Ana Jofre, Lan-Xi Dong
Ha Phuong Vu, Steve Szigeti and Sara Diamond
OCAD University
100 McCaul Street
Toronto, Ontario M5T 1W1, Canada

(Communicated by Aijun An)

Abstract. We present a method by which to convert a large corpus of website traffic data into interactive and practical taste graph visualizations. The website traffic data lists individual visitors' level of interest in specific pages across the website; it is a tripartite list consisting of anonymized visitor ID, webpage ID, and a score that quantifies interest level. Taste graph visualizations reveal psychological profiles by revealing connections between consumer tastes; for example, an individual with a taste for A may be also have a taste for B. We describe here the method by which we map the web traffic data into a form that can be displayed as interactive taste graphs, and we describe design strategies for communicating the revealed information. In the context of the publishing industry, this interactive visualization is a tool that renders the large corpus of website traffic data into a form that is actionable for marketers and advertising professionals. It could equally be used as a method to personalize services in the domains of government services, education or health and wellness.

1. **Background and introduction.** Taste graphs represent a means by which to chart, identify and compare relationships between consumer preferences [6, 7, 9, 5, 3]. Given a consumer's known taste (such as a preference for a particular brand), a taste graph identifies and visualizes other tastes and interests that a consumer is likely to hold. For example, Pearman [7] has identified that people who drive a Prius tend to also exhibit a great interest in computers and digital technology. One important application of the taste graph is to help guide targeted advertising. Taste graphs also drive the algorithms behind various recommendation engines, such as those used on Amazon [14].

Taste graphs that are designed as visualizations for the marketing sector have typically relied on consumer surveys [6, 9, 5, 3, 7] to chart relationships between consumer preferences. The work presented in this paper is not based on consumer survey data. Instead, we propose visualizations based on website traffic data, which has some advantage over consumer survey data. Surveys must be carefully designed to optimize response rates [13], and to avoid systematic errors such as social desirability bias [4]. Alternatively, collecting website traffic data can be less labour intensive than conducting surveys. On the other hand, the disadvantage of website

2010 *Mathematics Subject Classification.* Primary: 00A66; Secondary: 76M27.
Key words and phrases. MapReduce, taste graphs, data visualization, website traffic data, preferences.

traffic data is that patterns and relationships between aggregates are not immediately evident in these linear logs. Some amount of computational and design work is required to aggregate the data and to find possible correlations between the aggregates. This paper describes a method for constructing user-friendly interactive taste graphs from website traffic data, and we propose visualization strategies for communicating the results.

The concept for this approach was proposed by our industrial partner, who seek practical visualizations that reveal taste affinities between the various website pages they maintain. Our goal is to add value to their data by converting it from an unordered tripartite list into interactive visualizations that reveal connections between website visitors' interests. In other words, to convert a list of individual behaviours into a visualization of aggregate behaviour. The questions our visualization addresses are:

- Are those who visited page x likely to be interested in page y?
- Are those who visited page x more interested or less interested in page y than the average?
- Are those who visited page x more or less interested in page y than those who visited page z?

This type of cross-correlational information is of significant value to advertisers seeking to strategically buy space within the website and to better understand consumer preference patterns. We note that the use of such taste graphs has potential applications beyond the advertising industry, wherever there is a need to associate an expressed interest in one thing with an interest in another.

2. **Description of the interface and visualizations.** The intended user for our visualization design (and the target customer for our industry partner), to whom we refer as *the user* throughout the rest of this paper, is a marketing industry professional seeking to buy advertising space on the website, and seeking to better understand relationships between consumer tastes.

The first step in producing the visualization is to aggregate the data in a way that is meaningful for the user. We define a group as the set of all website visitors that have visited a particular webpage that is a component of the publisher's offerings. For example, all the people who have visited the publisher's theatre section form a group. The user can then ask questions like 'are people that are interested in theatre also interested in reading restaurant reviews?'

The goals of the visualization are to clearly convey the relationships between aggregates, and to facilitate the exploration of these relationships. The aim was to design an intuitive tool that allows users to:

- define a target group based on an interest (for example, people interested in A),
- display the target group's interests (i.e. show the other interests that people interested in A have)
- visualize how the target group's interests compare to others' interests (for example, are those interested in A more or less interested in B than the average? Are they more or less interested in B than those interested in C?).

When presented with the visualization, the user is prompted to select a target group. Once selected, the visualization displays the group's interest in all other pages on the site. The user can then select one or more specific interests for further exploration. To further explore a specific interest, the visualization displays

comparisons, allowing the user to see how much more or less interest the selected group has in a given page compared to a reference group. The reference group, by default, is the 'general public', which comprises all the website visitors in the entire website. Alternatively, the reference group can be another user-selected group.

Our data consists of a tripartite list in CSV format consisting of: visitor ID, web page ID, and the visitor's page score. The visitor's page score is what we use to measure of the degree of interest the visitor has in that page. We aggregate the scores with a simple average because it is an easy-to-understand statistical measure, and our target users may not be familiar with other statistical measures of aggregates. So, to determine the degree of interest that a group has in a specific page, we take an average of the page score over all the individuals within that group.

2.1. **Case study - walking through the interface.** The user's path through the interface, illustrated in figures 1 to 8, is as follows.

1. Select a group of interest, where groups are identified by visited page
2. Identify group's interests. View this group's average score for other pages they visited.
3. Comparisons
   (a) Compare group's interests with interests of the general public
   (b) Compare group's interests with interests of another group.

Figure 1 shows the initial landing page for the user upon opening our tool. We start by prompting the user to select a target group, which is defined by an interest. The data structure is of the form: Category/Interest, and we use this structure to avoid overloading the user with information. The opening page only shows the user a selection of categories. Once a category is selected, the user is then prompted to select an interest within that category on the main page of the interface, as shown in figure 2.

The display on the interface's main page divides the data into two parts: on the left, there is the target group, and on the right, there is their tastes. This strong visual separation was necessary to communicate that we are looking at a group of people on one side and their corresponding interests on the other. In informal pilot testing, we found that, because the groups are defined by a taste, users became easily confused between groups and tastes without an emphatic visual reminder. Spatial separation is an effective, and cognitively efficient, means to visually convey separate categories [11].

On the left side, the user is prompted to define a group by selecting an interest within the selected category. Each interest (or webpage) is inscribed within a bubble whose size is proportional to the total number of people who have visited that page. These size variations are intended to give the user a sense of the overall popularity of each of the interests in view. On the right side, the taste display on the right remains blank until the user defines a group, as shown in figure 2. In subsequent versions of this visualizations, we have also coded the data with shading, where the shading is proportional to the average score (averaged over all users in the data), and darker shades correspond to higher interest scores.

In figure 3, the user selects the circle labeled *American Food* in the group display on the left side. This means that the user has defined their target group as the website visitors that have looked at (shown an interest in) *American food* recipes.

Once a group has been defined, the group's interests populates the taste display on the right side, as shown in figure 3. Again, each interest is inscribed within a

bubble whose size visually maps the data. The area of these bubbles in the taste display is proportional to the number of people belonging to the target group who have visited this page. This is intended to give users an overall view of the levels of interest the group has in each page. In subsequent versions of this visualization, we also used shading in addition to size, where the size remains proportional to the number of visitors and the shading is proportional to target group's average score for this page (averaged only over the target group).

If the user wants to look at interests within other categories, s/he can use the upper right menu to navigate to a different category. Similarly, if the user wants to redefine their target group with a different interest category, the user can use the upper left menu to navigate to a different category. Figure 4 shows an example of the result when a user navigates to the *Hobbies and Interests* category.

When the user selects a taste from the display on the right side, a bar graph visualization pops up so that the user can start making comparisons, as shown in figures 5-8. In figures 5 and 7, the bar graphs are a direct visual map of the group's average interest score in the selected pages (which are the interest scores averaged only over group members), as well as a reference average interest score. In the case of figure 5, the reference group is the default 'general public', which comprises all individuals in the data set, and the reference interest score is a simple average over all individual interest scores.

We design the visualization interface such that the user has the ability to define a reference group in the same way they defined a target group. The user defines a reference group by simply selecting a second interest within the *group* display on the left. In the example shown in figure 7, the user has defined the reference group as the website visitors that have looked at *Asian food* recipes.

The scores listed for each interest in figures 5 and 7 are a proprietary measure that weigh different website engagement metrics, including number of visits and time spent. This proprietary score may not be an intuitive measure of interest for users, therefore we provide alternate strategies for comparison, and provide the means for the user to view their chosen group's interests mapped relative to the reference group.

In figures 6 and 8, each of the bars represents a comparison in the degree of interest between the selected group and the reference group; the bars visually map the factor by which the selected group is more or less interested in the page than the reference group. The length and orientation of the bars indicate the factor by which this selected group is more or less interested in the listed webpages than the general public. A bar to the right of the centreline indicates that the selected group is more interested in a particular page than the reference group by a factor proportional to the length of the bar, while a bar to the left of the centreline indicates that the selected group is less interested in a particular page than the reference group by a factor proportional to the length of the bar. In figure 6, the reference group is the default 'general public', while in figure 8, the reference group is defined by the user as the website visitors that have looked at *Asian food* recipes.

3. **Visualization strategy and rationale.** Our visualization, in accordance with good practice, offers the user several views of the data that provide different levels of detail [8]. The overview visually establishes two conceptual categories: on the left, a target group (who the data is about), and on the right, a list of interests/tastes (what the data is about).

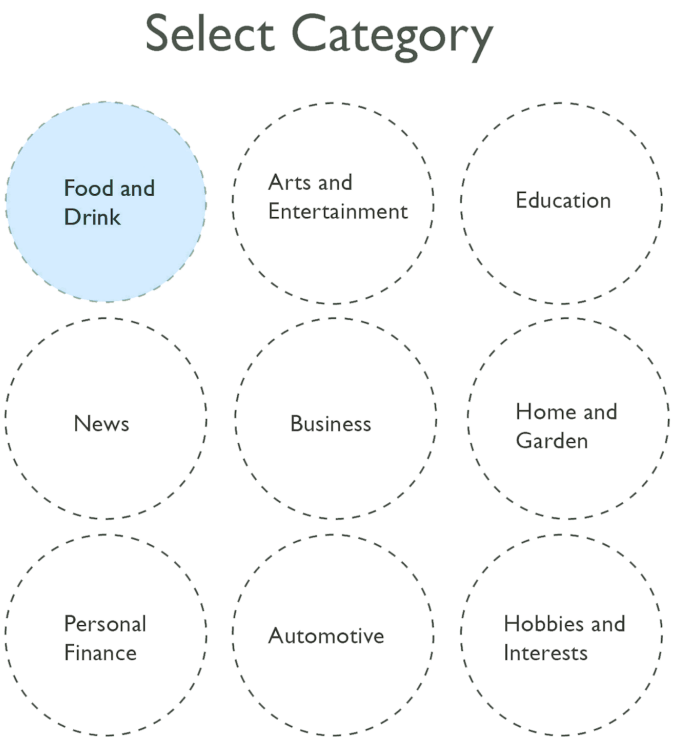

FIGURE 1. This is the first page the user encounters in our interactive visualization. The user is prompted to select a category. Categories are a means to filter the data so the user is not visually overwhelmed.

As the user walks through the visualization, different information and levels of detail about the information is revealed to provide the user with helpful overviews at every level [12]. The user starts with defining a target group by selecting a taste from the *group* display on the left side. The group display on the left side presents all the tastes within the user's selected category inscribed within circles whose area is proportional to the total number of visitors to that page. In later versions of the visualization, the shading of the circle denotes the overall level of interest in each taste (in other words, the average score for each page over all visitors). This gives the user a first level overview, and gives him/her a sense of how much general interest there is in each page. Once the user makes a selection to define a group of interest, the *taste* display on the right side reveals tastes inscribed within circles whose area is proportional to the total number of visitors to that page that belong to the selected group. In later versions of the visualization, the shading of the circle denotes the selected group's level of interest in each taste (in other words the average score for each page over the selected group only). This gives the user a second level overview, giving him/her a sense of how much interest their selected group has in each taste. Once the user selects one or more tastes, s/he can start looking at detailed comparisons in levels of interests between the selected group and

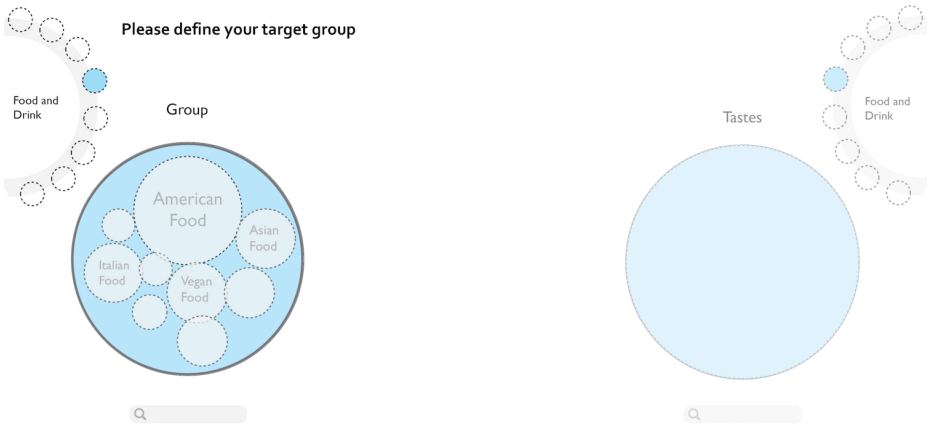

FIGURE 2. Once a category is selected, the user enters the main visualization view and is prompted to define a target group. The menu on the upper far left allows the user to navigate to a different category if s/he wants to browse through all the interests (website pages). A group is defined as all individuals with a common interest (or individuals who have visited a common page); the user choses the interest from a display on the left side of the visualization. The size of the bubbles in the *group* display on the left is proportional to the total number of visitors for each of the pages. The taste display on the right remains blank until the user has defined a group.

a reference group. We chose to use bar graph visualizations because they are most effective at quickly conveying quantitative comparisons between nominal categories [1].

We provide the viewer with two bar graph views. One displays the 'raw' scores for each selected taste, which, as explained previously is the page score averaged over all individuals within the selected group (or within the reference group). The other bar graph view displays a comparison: it visualizes the factor by which the selected group is more or less interested in a given taste than the reference group.

To visualize the factor by which one group is more (or less) interested in a taste than another, we start by taking the ratio of the selected group's page score to the ratio of the reference group's page score. When the ratio is greater than one, it means that the *selected group* is more interested in a given page than the reference group. Conversely, when the ratio is less than one, it means that the *reference group* is more interested in a given page than the selected group. We are careful to explicitly convey this information visually by drawing the bar on opposite sides of the centreline for each of these two cases. This way, the user can glance at the graph and quickly surmise which group is more interested in each page.

FIGURE 3. Here the user selects *American Food* in from the *group* display on the left. This defines the target group as all the individuals who have looked at (shown an interest in) *American Food* recipes in the website. Once the target group has been defined, the taste display on the right is populated. The size of the bubbles in the *taste* display is proportional to the number of target group members that have visited each of the pages. Above the *taste* display, there is a prompt for the user to select one or more tastes.

FIGURE 4. The user can use the menu on the far right to navigate to a different category. Here, the user has navigated the taste category from *Food and Drink* in figure 3 to *Hobbies and Interests* shown here. The user is still prompted to select one or more tastes.

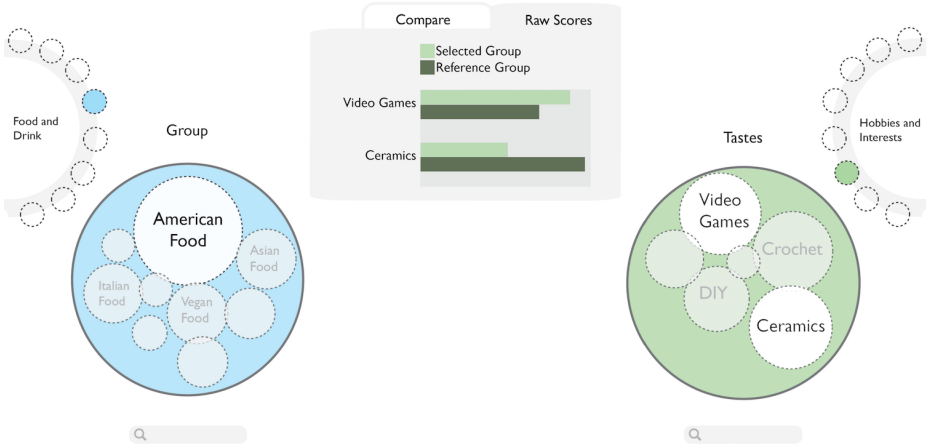

FIGURE 5. Here, there user has selected *Video Games* and *Ceramics*. This opens a window that allows the user to compare the target group's degree of interest in the selected tastes. The window has two tabs. The *Raw Scores* tab, shown here, displays the selected group's interest score for Video Games and Ceramics, as well as the reference group's interest scores. The default reference group is *general public*, which comprises all individuals in the data set.

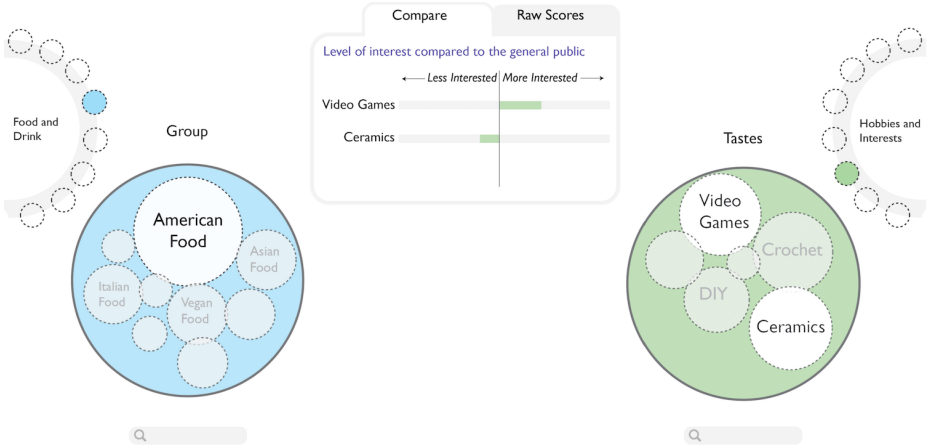

FIGURE 6. The *Compare* tab on the detail window visualizes the extent to which the selected group is more or less interested than the reference group in Video Games or Ceramics. Here the reference group is the default *general public*, which comprises all individuals in the data set.

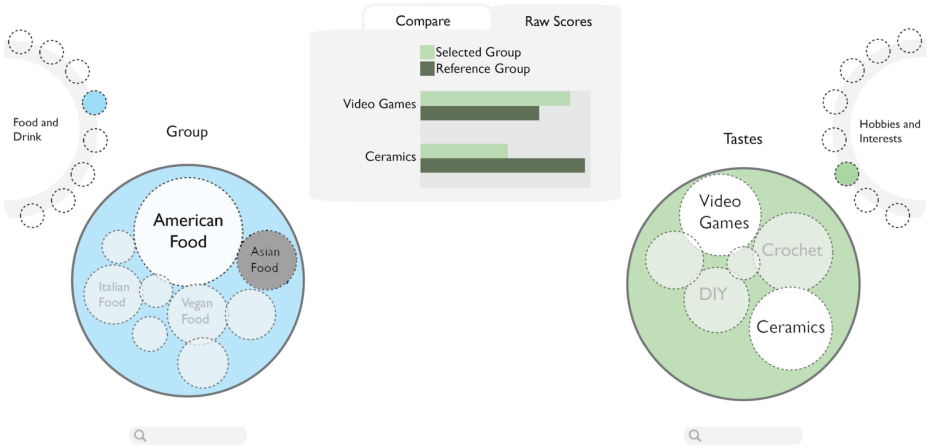

FIGURE 7. If the user wants the reference group to be something other than the 'general public', then s/he can define a reference group by selecting another interest in the *group* display on the left. Here the user selects *Asian Food* from the display. This defines the reference group as all the individuals who have looked at (shown an interest in) *Asian Food* recipes in the website. This view shows the detail window open to the *Raw Scores* tab displaying each group's interest scores in the selected tastes.

FIGURE 8. This view shows the detail window open to the *Compare* tab, where the user can directly compare how much more or less interest the target group (*American Food* recipe readers) has, relative to the reference group (*Asian Food* recipe readers) in Video Games and Ceramics.

In the case that, for a given page, the selected group's average score (GROUP) divided by the reference group's average score (REF) is greater than one, in other words if

$$R = (GROUP)/(REF) > 1, \tag{1}$$

then this means that the selected group is R times more interested in the page than the reference group. So we draw a bar that is R units long to the right of the centreline, on the *more interested* side of the divide.

Conversely, in the case that, for a given page, the selected group's average score divided by the reference group's average score is less than one, in other words if

$$R = (GROUP)/(REF) < 1, \tag{2}$$

then this means that the selected group has less interest in the page than the reference group. Therefore, in the case that R<1, we draw a bar that is 1/R units long to the left of the centreline, on the *less interested* side of the divide.

The use of bar length as a visual variable allows for clear visual comparison of the degree of interest between the listed pages. Bar graphs have the advantage that they can accommodate the visualization of an additional dimension by means of grouping [1, 11], and such flexibility is important as we develop our design.

4. **Programming methods: Preparing and aggregating the data for visualization.** Our objective is to aggregate the data in a way that is meaningful for the user. To develop our algorithm, we start with a medium-sized data set, consisting of 76 million instances of navigation through one of 1600 webpages, provided by our industry partner. The data is on a single csv file, and consists of three columns, one for visitor ID, one for visited page, and one for the score that quantifies the interest level in that page. To aggregate the data, we define a group as the set of all website visitors that have visited a particular webpage. For example, all the people who have visited the publisher's *American Food recipes* section form a group. Given a group, we track all the other webpages they have visited, and average over the interest scores for each of the other pages they visit. This allows our target user to ask questions like 'are people that are interested in cooking American food also interested in reading the latest news on video games?'

To process the data, we define one group for each of the visited webpages (for example, group x is the set of all who visited page x), plus one reference group that comprises all website visitors, for a total of 1601 groups. Then, for each of the groups, we do the following: 1) Extract the relevant subset of website visitors from the original data set. 2) Aggregate the number of visitors to each page, and average over the interest score for each page. The resulting output, after processing, is a list for each group consisting of: the pages visited (column 1), the number of group members who visited that page (column 2), and the average interest score of group members who visited that page (column 3). The processed data is much smaller in size than the original data, and it reveals aggregate behaviours and correlations that can then be visualized.

We analyzed a medium-sized data set consisting of 76 million entries, each an instance of someone visiting one of 1600 pages, with a total size of 10 GB. We put the data into a Python SQLite database, and run a query for each group, where (again) a group is defined as the list of all visitors who saw a specific page and the query extracts the group's level of interest in other pages. We reduced the

data to a set of 1601 lists, one for each group, with each list specifying the visited page (column A), the number of visitors to that page (column B), and the averaged interest score over all visitors to that page (column C). The total size of the reduced data (of all the combined lists) is about 10 MB, which easily fits onto the backend of a browser to support a visualization. Currently, our data is in CSV format and our visualizations are being developed in JavaScript with the D3 library [2].

5. **Discussion, future work and summary.** Currently, we store our data on our server in simple csv file format. However, in the context of this project, we will have to adapt our visualization framework to a different, more sophisticated back-end architecture to support the data we want to show.

The size of the processed data depends only the total number of webpages being tracked, not on the total number of visitors, so this method is in principle scalable. Increasing the size of the raw data set to 1 billion entries increases the statistical precision of the revealed correlations, but the processed data will remain the same size (as long as the number of webpages remain the same). The time to process the data, however, will increase.

As more website visitors are tracked, the data gets larger, and the computation gets slower. It took about 30 days to produce our preliminary results from 76 million entries on one of our lab computers. Typically, however, web-traffic data is much bigger than 76 million records, and we would like to extend our capabilities to be able to process much larger, more realistic, data sets. To this end, we will need a scalable data analytics platform that will allow us to develop algorithms to increase the efficiency of this calculation. Eventually, we want to be able to process and display the data real-time.

We are considering several changes in our visualization strategy as this work evolves. One thing that we have considered doing differently is the way in which a target group is defined. In our current procedure, we define a group as the set of all the website visitors that have visited a particular webpage. We note that another way to define a group is to select visitors whose interest in a particular page is above a threshold level of interest for that page. This increases the specificity of the group, and may be of interest to advertisers targeting a narrower group. Another change that we are considering is to find a way to display a distribution of interest scores, rather than just an average interest score. This latter change will require more computational storage and power than we currently have access to.

While the feedback we have received from our industry partners has been constructive towards defining our design, we have yet to run formal user evaluations on the design concepts presented here. We are currently in the process of developing a high-fidelity prototype with a subset of data for user testing.

In summary, we have reduced website traffic data from a long unordered tripartite list and created, without the use of consumer surveys, meaningful, applied real-world visualizations that directly reveal the level of interest a target group has in specific pages within the website.

REFERENCES

[1] A. Cairo, *The Functional Art: An Introduction to Information Graphics and Visualization*, New Riders, Berkeley, CA, 2013.
[2] M. Bostock, *D3.js - Data-Driven Documents*, Available from: https://d3js.org/.
[3] L. Kozlowski, Gravity can graph your internet clicks for new customer snapshots, http://www.forbes.com/sites/lorikozlowski/2013/11/13/

brand-graphs-a-new-snapshot-of-consumers/#7adbf713a2d1, *forbes.com*, 11.13.2013. Web 6.1.2016.

[4] I. Krumpal, Determinants of social desirability bias in sensitive surveys: A literature review, *Qual. Quant.*, **47** (2013), 2025–2047.

[5] V. Kumar, Building a Taste Graph: The basic principles, http://bigdata-madesimple.com/category/tech-and-tools/analytics/, *bigdata-madesimple.com*, 12.23.2014. Web 6.1.2016.

[6] S. Pearman, Delicious interest graphs: Taco bell and whole foods, http://www.gravity.com/blog/delicious-interest-graphs-taco-bell-and-whole-foods/, *Gravity.com*, 8.12.2013. Web 6.1.2016.

[7] S. Pearman, What Your Electric Car Says About You, http://www.gravity.com/blog/what-your-electric-car-says-about-you/, *Gravity.com*, 7.31.2013. Web 6.15.2016.

[8] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, *IEEE Symposium on Visual Languages Proceedings*, (1996), 336–343.

[9] A. Taylor, Snow Fight: Skiing versus Snowboarding, http://www.gravity.com/blog/snow-fight-skiing-versus-snowboarding/, *Gravity.com*, 2.17.2015. Web 6.1.2016.

[10] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd edition, Graphics Press, Cheshire, Conn., 2001.

[11] C. Ware, *Information Visualization: Perception for Design*, Elsevier, Waltham, MA, 2013.

[12] N. Yau, *Visualize This*, John Wiley & Sons, Indianapolis, Indiana, 2011.

[13] J. Yu and H. Cooper, A quantitative review of research design effects on response rates to questionnaires, *J. Mark. Res.*, **20** (1983), 36–44.

[14] T. Zhou, J. Ren, M. Medo and Y.-C. Zhang, Bipartite network projection and personal recommendation, *Phys. Rev. E*, **76** (2007), 046115.

E-mail address: jofrea@sunyit.edu
E-mail address: ld13jh@student.ocadu.ca
E-mail address: hvu@faculty.ocadu.ca
E-mail address: sszigeti@ocadu.ca
E-mail address: sdiamond@ocadu.ca