# VizClick
## *Visualizing Clickstream Data*

Rajat Kateja[1,3,*], Amerineni Rohith[1,3,*], Piyush Kumar[2,3*] and Ritwik Sinha[3]
[1]*Indian Institute of Technology Guwahati, Guwahati, India*
[2]*Indian Institute of Technology Bombay, Mumbai, India*
[3]*Adobe Research, Bangalore, India*

Keywords: Clickstream, Visualization, Association, Geo-spatial, Site Topography, Big Data.

Abstract: Clickstream data is ubiquitous in today's web-connected world. Such data displays the salient features of big data, that is, volume, velocity and variety. As with any big data, visualizations can play a central role in making sense and generating hypotheses from such data. In this paper, we present a systematic approach of visualizing clickstream data. There are three basic questions we aim to address. First, we explore the inter-dependence between the large number of dimensions that are measured in clickstream data. Next, we analyze spatial aspects of data collected in web-analytics. Finally, the web designers might be interested in getting a deeper understanding of the website's topography and how browsers are interacting with it. Our approach is designed for business analysts, web designers and marketers; and helps them draw actionable insights in the management and refinement of large websites.

## 1 INTRODUCTION

A clickstream is defined as the series of mouse clicks made by the user of a website. For websites, clickstreams serve as a source of highly valuable information. In particular, it tells the website owners about their users, and what they are interacting with. This information helps in aiding marketing decisions and help in providing a personalized experience. While the information contained in clickstream data is beyond dispute, it comes with many challenges. These challenges stem mostly from the fact that clickstream data displays all aspects of big data (McAfee et al., 2012). Namely, clickstream data is high in volume and velocity, with one day's worth of such data amount to many tens or hundreds of gigabytes for any major website. Further, clickstream data is large in variety and is inherently heterogeneous (Wei et al., 2012).

Information visualization may be defined as a visual representation of data that helps in generating hypotheses and making inferences. The history of data visualization goes a long way back (Tufte and Graves-Morris, 1983). The second half of the 20$^{th}$ century saw the formal development of the grammar of graphics (Cleveland, 1993). The last two decades have seen an explosion in the development of platforms for the development of professional quality information visualizations, R (R Core Team, 2013) being one of the more popular platforms. While R has significant capabilities to generate high quality graphics (Sarkar, 2008; Wickham, 2009); there are limitations in creating highly interactive visualizations. Interactive visualizations are important for big data because they give the user the ability to hierarchically explore the data. There have been some effort to create platforms for interactive visualizations (Swayne et al., 2003), but, much of these have been overshadowed lately by JavaScript based graphing libraries [2]. We see two advantages of these, first, one can have customizable levels of interactivity, and second, they can be viewed on a browser, making them easier to share. Among these libraries, D3 (Bostock et al., 2011) has received considerable interest and as a result is a mature framework for interactive visualizations. Much of the visualizations created in this paper use and build on D3.

*First three authors have equal contribution. We would also like to acknowledge Sandeep Zechariah George K (sgeorge@adobe.com) for his continuous support during the project and especially for his guidance with D3.

[2]http://socialcompare.com/en/comparison/javascript-graphs-and-charts-libraries

Kateja R., Rohith A., Kumar P. and Sinha R..
VizClick - Visualizing Clickstream Data.
DOI: 10.5220/0004687102470255
In *Proceedings of the 5th International Conference on Information Visualization Theory and Applications* (IVAPP-2014), pages 247-255
ISBN: 978-989-758-005-5

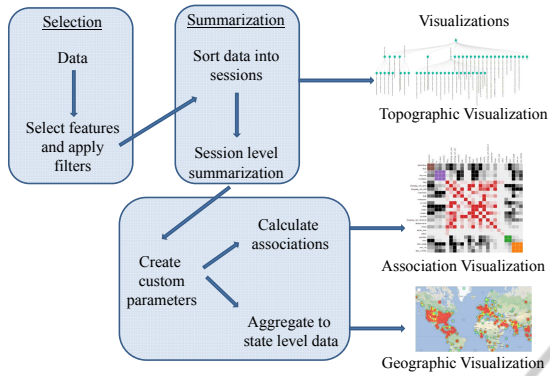


Figure 1: A schematic representation of the work flow. We start with filtering the raw data, which is then sorted and summarized into sessions. This sorted data is used to create transition matrices for visualizing the site topography. The session level data is further processed to create visualizations of association and at the geographic level. Finally, the entire solution was packaged as a website.

In this paper, we present a framework for answering important questions from clickstream data using visualizations. Adobe© Analytics [3] provides marketers the most comprehensive set of tools to measure and track all aspects of usage of an organization's website. We built and tested our framework on the clickstream data of www.adobe.com collected using Adobe© Analytics. We deal with three aspects of clickstream data.

First, as is evident, many parameters of interest to the marketer are collected in clickstream data. For example, referring domain, time spent on the website, number of clicks in a single session, visits to different site sections, and events on the website (like purchase or downloads). Some of these large number of parameters may be associated with each other, the presence or absence of such associations is informative to the marketer. Given the large size of data and the number of features, these associations are not apparent. In this paper, we evaluated a number of approaches to measure association between two features in the context of clickstream data. Given a measure of association, an equally challenging aspect is to present these pairwise associations among a large set of parameters in a way that enables finding patterns.

Next, many website owners (particularly e-commerce websites) are interested in analyzing geospatial aspects of the clickstream data. To address questions like, "In which regions are our products in demand?"; "Where are we doing poorly?"; "If we had some marketing budget, what regions should we concentrate on?" We analyzed clickstreams and summarized it at some geographical level. We used a continuous area Cartogram (Dougenik et al., 1985) as a means to present this data and enable information extraction. We next presented geographic data in a manner that helps detect deviations from expectation, thus aiding anomaly detection.

Finally, we contrast the site design with how users interact with it. The web designer has a notion of relationships between different content on the website, however, users may have a very different view of these relationships. The clickstream data tell us how users see the structure of the website. Such comparisons provide meaningful insights to web designers on how to perform minor or major reorganizations of their website. There have been several studies on site topography (Lee et al., 2001; Brainerd and Becker, 2001; Ferreira de Oliveira and Levkowitz, 2003), but they do not propose interactive visualizations and do not allow for analysis at higher levels of granularity (up to the level of individual pages). We created visualizations, which incorporate custom clustering, so that the website owner can focus on relevant areas and make deductions.

The rest of the paper is organized as follows. The second section explains the data used as well as the pre-processing necessary. It also describes the methods devised to address the three broad goals of the paper. The third section is dedicated to results which show how making inferences is made easier with the help of such visualizations. The last section deals with conclusions and future work.

## 2 DATA AND METHODS

The process of going from the raw clickstreams to the final visualizations from which a user can extract actionable insights is a long one (Figure 1). We had the entire clickstream data of all traffic to www.adobe.com for three days. Each click had measurements on 250 features. The first step in our work involved filtering columns and rows. Many of the 250 features were missing (either partially or completely). Using the description of the data, we restricted our analysis to 70 fields which had relevant information. Some of the fields are: visitor IDs, visit number, hit time, referrer, browser, resolution, geographic region, IP, page section. Also, given our interest is spatial visualizations, we restricted our data to only those clickstreams that originated in the USA, to ensure homogeneous data. This filtering led to each day having about 10 million clicks and about a third as many sessions.

For the ease of computation at a later stage, we

[3] http://www.adobe.com/in/solutions/digital-analytics/marketing-reports-analytics.html

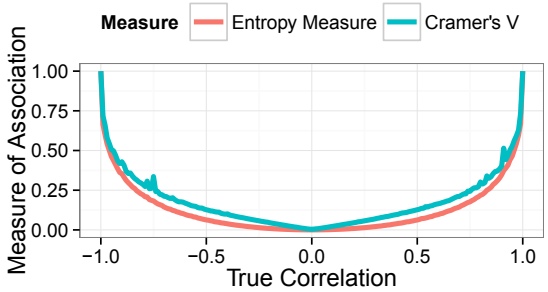


Figure 2: Comparison between Cramer's V and Entropy Measure.

sorted this data by session. This sorted data was used to compute the transition matrices needed for the site topography visualizations. For visualizing associations we summarized data at the session level. The data was further aggregated on the basis of geographic location (States of the USA) for the geo-spatial visualizations. Finally the entire solution was packaged as a webpage with sections dedicated to each type of analysis, where one could select parameters (like time of analysis or features to consider). We next describe the visualizations in more detail.

## 2.1 Association

A number of parameters of interest are generated in clickstream data. The first question to address in any data analysis is to identify the inherent structure in the data. In other words, is there a suggestion of relationship between different measured dimensions. The best measure of association between continuous variables is the correlation coefficient (Kutner et al., 2004). However, in clickstream data, a number of parameters of much interest are categorical in nature (Agresti, 2002), that is, there is no inherent order in the values (for example, operating system, referring domain and purchase flag). Hence, we want a measure of the strength of association between two variables when either of them can be a continuous or categorical variable.

As a solution, we binned the continuous variables to convert them to categorical variables. Such an approach was necessary because categorical variables cannot be converted to continuous variables. Now that we had decided that all the variables will be categorical in nature, we still needed to decide upon the best measure of association between two categorical variables. While there is considerable work on defining measures of association between two categorical variables (Press, 1992), we concentrated on the following as they are better accepted in practical usage[4]

[4]Please note here that the $p$-value of $\chi^2$ test of independence of the contingency table is a measure of the deviation from the null hypothesis (independence) and will converge to 0 as the size of data increases. Hence, this measure is not usable for our problem.

: C coefficient, Cramer's V and Entropy based association measure.

### 2.1.1 Choice of Measure of Association

Of the three measures considered, the C Coefficient does not take into account the number of bins of the categorical variable, and hence could not be used to compare the association between pairs of variables unless they have equal number of bins. Hence this measure was not appropriate for our situation. So we were left with two option, Cramer's V and Entropy based measures. We conducted experiments with these to see which one suits our needs the best. We generated bivariate normal data with a known correlation structure (referred to as true correlation in Figure 2). Then we binned the random numbers generated to convert them into categorical variables. Next, we calculated their association using these two measures. The results have been plotted in Figure 2. First, notice that both measures generally increase with increasing correlation coefficient, which is a desirable property. Second, both measures are bounded between $[0,1]$, which is important for comparability. Finally, both measures, though increasing, do not have an linear relationship. Although both measures behave similarly, Cramer's V lacks the smoothness that one would desire in a measure of association. Hence, we decided to use the entropy based association measure in our work.

### 2.1.2 Entropy based Association

Here is a brief description of the entropy based association measure. Given two categorical variables $X$ and $Y$, with $m$ and $n$ possible values, respectively, let $N_{ij}$ be the number of observation which have the $i^{\text{th}}$ value of $X$ and $j^{\text{th}}$ value of $Y$. Similarly, let $N_{i.}$ be the total number of observations having the $i^{\text{th}}$ value of $X$, $N_{.j}$ be the total number of observations having $j^{\text{th}}$ value of $Y$ and, let $N$ be the total number of observations. Mathematically,

$$N_{i.} = \sum_j N_{ij}, \quad N_{.j} = \sum_i N_{ij}, \quad N = \sum_{i,j} N_{ij}$$

Next, define

$$p_{ij} = \frac{N_{ij}}{N}, \quad p_{i.} = \frac{N_{i.}}{N}, \quad p_{.j} = \frac{N_{.j}}{N}.$$

The bivariate and univariate entropy measures are defined as follows,

$$H(X,Y) = -\sum_{i,j} p_{ij} \ln p_{ij},$$
$$H(X) = -\sum_{i} p_{i.} \ln p_{i.},$$
$$H(Y) = -\sum_{j} p_{.j} \ln p_{.j}.$$

Finally, the entropy based association measure is,

$$U(X,Y) = 2[\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}].$$

#### 2.1.3 Practical Concerns and Visualization Techniques

Having decided on a measure of association, we had to decide on the optimal number of bins when converting a continuous variable into a categorical one. A number of approaches have been proposed to address this (Sturges, 1926; Freedman and Diaconis, 1981). We used the Sturges method, as the other approaches were suggesting too many bins, leading to high cardinality of the resulting categorical variable.

Once all the variables were converted to categorical variables, we still needed to ensure that the cardinality of these variables was not too large. This was important for the validity of the association measures, which behave badly when some marginals of the contingency table are sparse. To reduce the number of possible values of a categorical variable, we pooled all less frequent ($\leq 0.1\%$ of the data) values into a group called "other".

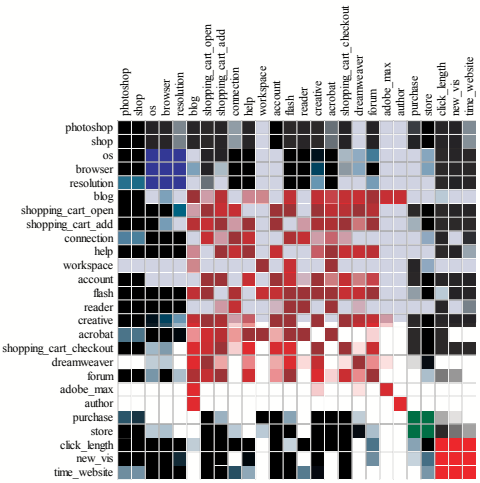


Figure 3: A heatmap representation of the association matrix. Each row and column represents some parameter of interest. The color intensity at some row and column denote the strength of this association. Each different color denotes the set of parameters that form a homogeneous cluster of highly dependent parameters.

After computing the associations based on the entropy measure, we considered a number of ways in which they can be visualized. Our first approach was to use an interactive heat map of the association matrix. Next, we represented the associations as a graph, with nodes being the parameters and edges representing the strength of association. We used the forced layout to display this graph. In addition to these, we used a chord layout to represent the association values between each pair of parameters. The thickness of the chord corresponding to the association value. We further added an animation to the chord diagram to represent associations changing over time. All the mentioned visualizations and their interactivity are explained in the Results section.

### 2.2 Geographic Visualizations

The most granular representation of a geographic location is the latitude and longitude. Unfortunately, this information is not available in our clickstream data. In our experiments, deriving this from the IP address is error prone. Hence, we decided to restrict our analysis to some higher level geographic entity. In our data, when restricting to clicks from the USA, the State of the visitor was consistently available. Hence, all our analysis in this section is restricted to analyzing data at the state level.

The first step to making geographic visualizations, is to summarize data at some geographic unit (in our case, the State). We used Cartograms (Dougenik et al., 1985) to visualize features by State. A Cartogram is powerful as it provides multiple encodings for representing information. While color is used as one encoding, it gives the additional ability to use rescaled area as a visual representation. This dual encoding is particularly useful for users with achromatic vision as they can infer the proportions based on the region's scaled area. However, this does require the user to have ready access to the original unscaled map, to observe differences. We start our visualization with a representation of the unaltered map and upon feature selection by the user, the map smoothly transitions into the Cartogram, representing a metric of interest.

Most features considered for association were considered here as well. The visualizations were performed for each of the three days and across all features. The transitions over features or days allows for contrasting differences. The results section presents some samples of this visualization and its interactivity.

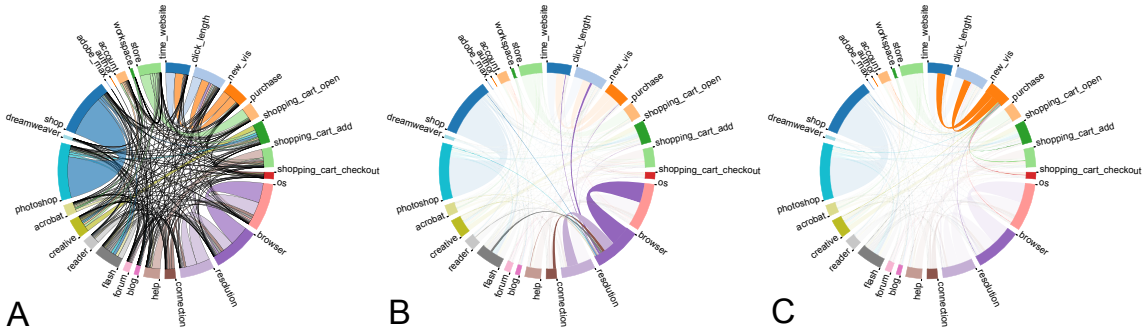


Figure 4: Figure A shows the default view of the chord diagram. Figure B shows the "browser" feature highlighted, it is clear that it is highly associated with "OS" and "resolution". Figure C has the "new visitor" highlighted and it is strongly associated with "time spent on website" and "average click length of session". This piqued our interest and upon further exploration, we found that new visitors tend to spend lesser time on the website and have fewer clicks.

While the Cartogram provides a succinct visualization of a particular metric, there are some situations where it might be misleading. One such situation is when the user wants to contrast the observations with some expectation. For example, one may expect that visits to www.adobe.com follow the distribution of internet population across States. Deviations from this expectation may be of concern. With this motivation, we created a visualization that highlights such deviations, using color as a visual encoding to represent deviations. Expectations other than internet population may be based on State Gross Domestic Product (GDP) or State Disposable Income, which may be of more interest to a marketer. The marketer may be concerned that a state with significant GDP is contributing little to sales. We refer to such infographics, which even out the metric based on some underlying factor as "normalized geographic visualization" in the rest of the paper.

## 2.3 Website Topography

Next we propose a series of visualizations that help the website owner understand how the users interact with the website to identify possible areas of improvement. The first step in doing this is the creation of a transition matrix based on the clickstreams; which stores how many times users have gone from one page to another. Alternatively, this information may be stored as a directed graph, where each node is a page, and the edge is the volume of traffic between them (in the given direction). In addition to the pages as nodes, we added two artificial nodes to represent entry to the website and exit from the website.

The first question to address for a website owner is to understand which pages lead to the highest bounce rate, that is, a high proportion of users arriving at the page are leaving the site altogether. To represent this, we used a bar graph, where the length of the bar represents the bounce rate. Additionally, we used color to represent the number of users arriving at the particular page, thus identifying the most important pages where the bounce rate is a worry.

Next, we delve more into how users interact with the website. Our initial attempt was at displaying the website as a graph with the edges representing transitions. Unfortunately, in our data, we had of the order of $10,000$ distinct pages. This meant that visualizing the raw graph was not feasible. The structure of this network is lost if we attempt to visualize only the more active nodes. To overcome this challenge, we had to cluster nodes in some manner. We explored two different approaches to do this.

First, we used the modularity based approach to detecting communities in large networks (Blondel et al., 2008). This algorithm, when given a graph, produces successive clustering which is based on the connectivity of the nodes in the graph. Successive clustering leads to a hierarchical tree, leaf nodes representing individual webpages and internal nodes representing clusters. This tree is then visualized using an interactive bubble plot. The bubble plot creates a bubble for each leaf node, and combines these to form larger bubbles representing a cluster, which may further group to form a higher level cluster. The size of a bubble is proportional to the total traffic on that page. We also name the cluster by identifying the major keywords in URLs for each cluster, where keywords are defined as sections of website to which webpages belongs. Another level of interactivity is added by providing links to the actual pages which upon clicking leads the user to the page.

The second way to cluster the webpages was based on the hierarchy in the site structure. Here we use the URL's of the webpage to identify parents to webpages and cluster based on these. This clustering is then visualized as a collapsible tree, where the user can expand only the web-sections she wants to

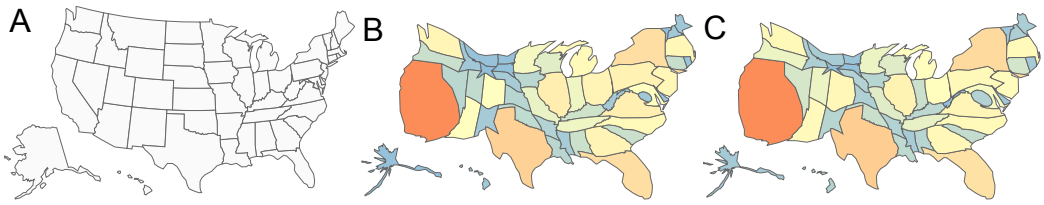


Figure 5: Figure A shows the original map of US with no distortion, the default view. Figure B shows the Cartogram for number of new visitors for each State. Figure C shows the number of purchases from each State. The inflation in size of Washington DC is clearly apparent from Figure B and C. Also note that California is blown up in both of them, possibly indicating a pattern that needs further exploration.

concentrate on. As the next step, to make the traffic patter apparent, we added interactivity. When a user hovers over any node in the tree, all its highest degree neighbors (either in-degree or out-degree) in the graph are highlighted (by increasing the node's radius proportional to the edge weight). This view provides an interactive view of the user behavior keeping the web designer's site structure intact.

In addition to the above, we considered another interesting question which can be answered from clickstream data. If the data suggests that a lot of users are navigating from page *A* to page *C*, via a set of pages $B_i$, then it may be interest to see if a link to page *C* can be provided on page *A*. We identified such pairs based on the transition matrix and visualized them as a bar chart.

## 3 RESULTS

In this section, we show the visualizations we have created and highlight some inferences that can be easily drawn from them.

### 3.1 Association

Figure 3 shows the heatmap of the association matrix. The association matrix was used as a distance matrix to create clusters of parameters, based on hierarchical clustering. For example, the largest cluster in the center has the sections of the website that are frequently visited together, such as, "shopping cart open" and "shopping cart add". Each cluster is shown with a color and cross cluster cells are shown in black. The color intensity represents the strength of association, higher associations having darker shades. As expected, each variable is perfectly associated with itself. An interesting aspect that can be easily identified from this view is the small cluster of three parameter "OS", "browser" and "resolution", which points to the high association between the three.

A graph is a natural representation of a distance matrix. Hence, we next created a graph with a forced layout. The nodes are the parameters and edge weight is the association between each pair of nodes. The graph (not shown in the paper) is, unfortunately, too cluttered to draw any significant inferences. This motivated us to explore some other visualization which presents the same information, with less clutter. Hence, we used a Chord diagram to represent the same information (Figure 4).

In the first view of this visualization, the arc length provided to each of the parameters is proportional to the sum of its association values with all other parameters. For example, "OS" has a higher arc length than "account", representing that "OS" is more associated with other parameters, whereas "account" is not related to any other parameter. Also clearly noticeable is the thick chord between "shop" and "photoshop", perhaps indicating that a lot of people looking to shop on Adobe's website are interested in Photoshop®. In part B (which can be achieved by hovering over a parameter), the focus shifts to "browser" and helps the user concentrate on this particular aspect. As noted from the heatmap as well, "browser" is highly associated with "resolution" and "OS". Finally in part C, the "new visitor" parameter is highlighted (which is a binary flag indicating whether the visitor is new or has visited before). Notice the high association between "new visitors" and the "time spent on the web-

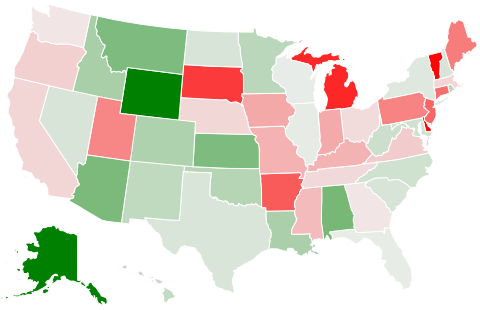

Figure 6: Geographic visualization of purchases normalized for the internet population in each state. California, which was highlighted in Figure 5, is now seen to be following expectations.

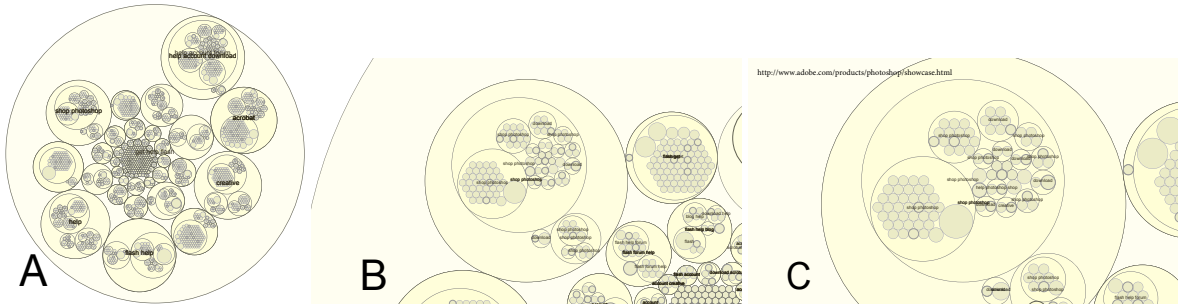


Figure 8: Figure A shows the entire hierarchy of the website based on modularity as successive bubbles. Next, Figure B shows one of the bubbles focused on, and the names of various other sub-bubbles are also visible which gives an idea of the prominent tags in those bubbles. Figure C shows the URLs that are visible upon hovering over an individual webpage. These bubbles also serve as links to move directly to those webpages.

site". This raises a hypothesis that may be tested to get more insights from the data. In such an exploration, we found that new visitors spend lesser time on the website than others. Since all the parameters are categorical in nature with no inherent ordering, the actual trend cannot be predicted based just on the association values.

## 3.2 Geographic Visualizations

Figure 5 provides views of the Cartogram. In part A, we see the original map of the USA with no scaling. In part B, the Cartogram is drawn according to the metric "new visit" (number of new visitors from each State). Part C shows the Cartogram "purchases". An interesting insight is that the size of Washington DC inflates considerably when purchases are considered. We provide the option to visualize various metrics like these and cross metric comparisons are facilitated by smooth transitions between Cartograms. Such information and comparisons provide actionable insights for marketers.

While the Cartogram provides an effective tool to use size as a visual encoding for a metric, there are important patterns that are not highlighted. For example, in all the cartograms we created, we see consistently that California is significantly larger than its original size, could this be because California has a large share of the internet users of the USA? The Cartogram fails to answer this question. This led to a need for visualizations based on normalizations, that help detect deviations from what is expected.

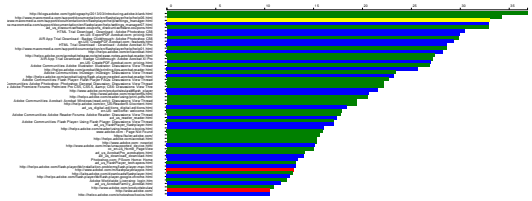

Figure 7: Pages with the highest bounce rates. The length denotes the proportion of bounce, and the color denotes the number of visitors. Red signifies the highest traffic pages, yellow is next, followed by blue and then green.

In the normalized geographic visualization presented in Figure 6, red implies negative deviation and green represents positive deviation from expected behavior, with light grey being the neutral color. As can be seen, California is very close to grey implying that the number of purchases from there are as expected according to population of internet users in that state. Alaska on the other hand is bright green, generating a finding that may warrant further exploration. Similarly, Michigan is deep red, and draws the marketer's attention as expected from such a visualization.

## 3.3 Website Topography

As mentioned earlier, before delving deep into website topography, we first started with some primary questions that a website owner may have. Figure 7 displays the pages with the highest bounce rates, the color coding further denoting which of these are the highest traffic pages.

We next clustered the webpages based on modularity and visualized them using the bubble plot. Figure 8 shows this visualization. Part A shows the default view of the visualization, with successive clusters represented as successive bubbles. The size of a node is proportional to the sum of the in-degree and out-degree. Part B displays the interactive capabilities of the visualization, here a particular bubble has been brought into focus by clicking on it. Bubbles have been named depending on the major sections of the website that are a part of the bubble's nodes. In part C, a particular webpage node has been highlighted which displays the URL of the webpage on the bottom of the visualization. The visualization interface also supports clicks on the node to directly move to

the webpage representing the node.

Next, we created a visualization based on the website's structure (Figure 9). As explained earlier, in this visualization, hovering on particular node leads to its neighboring nodes being highlighted (increases in size). This increase is proportional to the edge weights between the two nodes. By doing so, we avoid unnecessary cluttering in representing the edges

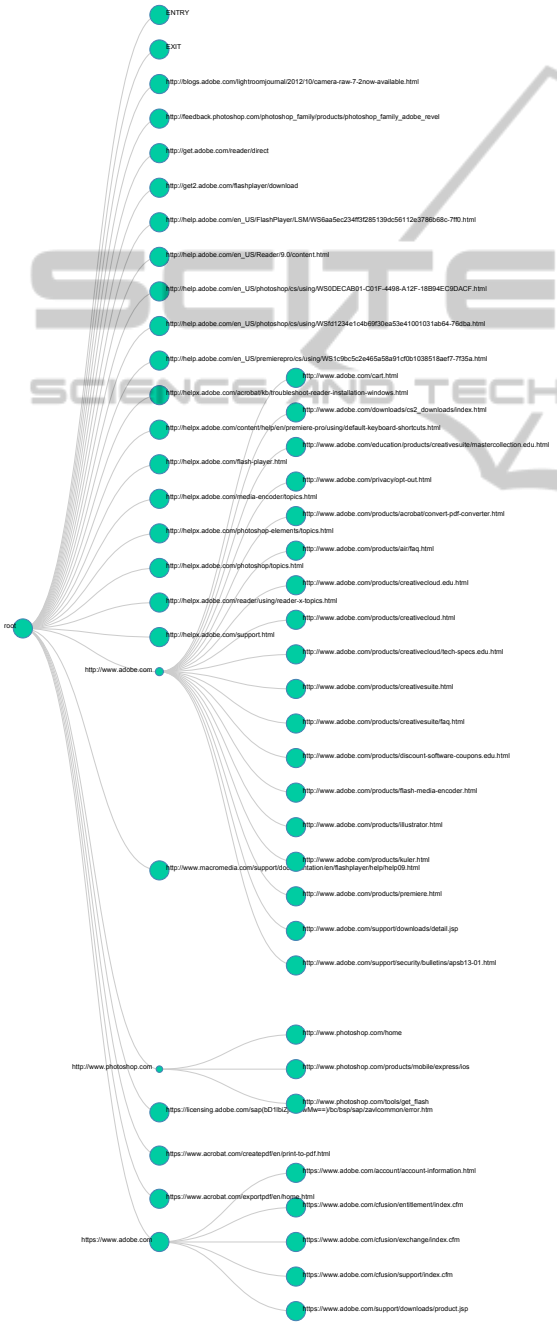


Figure 9: Site Topography. Hovering over a node increases the size of its neighbors (based on in-degree or out-degree).

but still provide access to the information. Also by examining the inter-cluster traffic, analysts can get an idea of cross cluster movements. If inter-cluster movements are highly significant, the web designer can consider redesigning their site structure, to provide a better user experience. Another use of this visualization is as follows: Suppose, there is a sudden drop in the number of users moving from page $A$ to page $B$. From the visualization presented, such sudden changes can be easily observed, and may point to a broken link or other anomalies.

Lastly, we identified suggested links for certain pages based on the traffic pattern. We visualized this as a histogram (not shown in the paper), with each row corresponding to a pair of webpages $(A,C)$. The pair is such that it might improve user experience if a link to page $C$ is added on page $A$.



## 4 CONCLUSIONS

We have compiled a collection of visualizations that can be used to understand three important aspects of clickstream data, namely, association between parameters, spatial aspects of clickstream data and site topography. We have shown that a number of interesting and useful insights can be generated based on the visualizations created. In addition, these visualizations serve as a way of generating hypotheses that can then be tested on the data.

The proposed visualizations can serve as the first step in making sense out of large web analytics data. Additionally, we have focused extensively on creating visualizations that are interactive. Given the size of clickstream data, it is important for the user to have the ability to dig down further into areas that might be of interest. Also, interactivity provides a way of visualizing large amounts of data without over powering the visual senses of the viewer. Our visualizations of the website topography are designed to achieve this goal. However, in the context of visualizing large graphs, there is scope for building in even higher levels of interactivity without losing out on the ability to communicate information.

## 5 ADDITIONAL RESOURCES

Please visit this link for a short video description of VizClick, highlighting some of its interactivity: http://youtu.be/fv9qrU0EZJE.

## REFERENCES


Agresti, A. (2002). *Categorical data analysis*, volume 359. John Wiley & Sons.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). $D^3$ data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309.

Brainerd, J. and Becker, B. G. (2001). Case study: E-commerce clickstream visualization. In *infovis*, pages 153–156.

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

Dougenik, J. A., Chrisman, N. R., and Niemeyer, D. R. (1985). An algorithm to construct continuous area cartograms. *The Professional Geographer*, 37(1):75–81.

Ferreira de Oliveira, M. C. and Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378–394.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L-2 theory. *Probability theory and related fields*, 57(4):453–476.

Kutner, M. H., Nachtsheim, C., Neter, J., et al. (2004). *Applied linear regression models*. McGraw-Hill New York.

Lee, J., Podlaseck, M., Schonberg, E., and Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. In *Applications of Data Mining to Electronic Commerce*, pages 59–84. Springer.

McAfee, A., Brynjolfsson, E., et al. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–66.

Press, W. H. (1992). *Numerical recipes in Fortran 77: the art of scientific computing*, volume 1. Cambridge university press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sarkar, D. (2008). *Lattice: multivariate data visualization with R*. Springer.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66.

Swayne, D. F., Lang, D. T., Buja, A., and Cook, D. (2003). Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444.

Tufte, E. R. and Graves-Morris, P. (1983). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Wei, J., Shen, Z., Sundaresan, N., and Ma, K.-L. (2012). Visual cluster exploration of web clickstream data. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 3–12. IEEE.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated.