Hypertext'14 Workshop

Data Visualization Workshop (DataWiz 2014)

Welcome to DataWiz 2014, the 1st workshop on Data Visualization, one of the satellite events of the 25th ACM Hypertext conference, that is a premium venue for high quality peer-reviewed research on hypertext theory, systems and applications.

DataWiz aims at bringing together an interdisciplinary audience (e.g. computer and cognitive scientists, designers, data journalists), in order to discuss tools, models and metaphors useful to understand and explain input or output data through advanced graphical user interfaces.

In fact, interacting with data by means of intriguing visual representations, that can eventually be accessed from the Web, is a fundamental approach to accurately present scientific findings in an appealing way. The aim is to leave to the observer (likely an expert of the domain) the task of exploring complex phenomena, without the pain of dealing with issues such as data complexity and overload of information. This workshop focuses on both scientific and information visualization, with the aim of discussing of best practices and innovative approaches.

12 papers out 17 submissions were accepted for 20 minutes presentations; the morning session will open with the invited speech of Everardo Reyes-Garcia, whose contribution is included in the proceedings as well.

All papers were initially assigned to be reviewed by three members of the PC. Final decision on borderline papers was made on meta-reviews written by the PC chairs. Thanks to their effort, the program includes lots of exciting ideas that we can't wait to share with you in our full-day workshop that will be held in Santiago, on September 1st.

We thank all participants of the workshop for their contributions and the organizers of the Hypertext 2014 conference for their support. We hope that you will find this program interesting and thought-provoking and that the workshop will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world. We are looking forward to a very exciting and interesting workshop.

Martina Deplano	*André Panisson*	*Giancarlo Ruffo*
University of Turin	ISI Foundation	University of Turin
Turin, Italy	Turin, Italy	Turin, Italy

Program
(accepted papers in authors' names alphabetic order)

Cazabet Remy and Takeda Hideaki, *A Visualization Platform For Exploring Cooperation*

Celestini Alessandro, Di Marco Antonio and Totano Giuseppe, *A Data Extraction and Visualization Framework for Information Retrieval Systems*

Ferrara Emilio, De Meo Pasquale, Catanese Salvatore and Fiumara Giacomo, *Visualizing criminal networks reconstructed from mobile phone records*

Graves Alvaro and Bustos-Jiménez Javier, *Towards Visual Overviews for Open Government Data*

Honaker James and D'Orazio Vito, *Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories*

Honorato Johanna, Cypriano Lucas, Goveia Fabio and Carreira Lia, *The color of the street: color as images visualization parameters of twitter pictures from Brazilians Manifestations of 2013*

Medeiros Jean Maicon, Regattieri Lorena and Malini Fabio Luiz, *The use of modularity algorithms as part of the conceptualization of the perspectival form in large networks*

Móro Róbert, Daráž Jakub and Bielikova Maria, *Visualization of Gaze Tracking Data for UX Testing on the Web*

Regattieri Lorena, Chartier Ryan, Windsor Jennifer and Rockwell Geoffrey, *TweetViz: Following Twitter Hashtags to Support Storytelling in Data journalism*

Regattieri Lorena, Goveia Fabio, Herkenhoff Gabriel and Malini Fabio Luiz, *MarcoCivil: Visualizing the Civil Rights Framework for the Internet in Brazil*

Reyes-Garcia Everardo, *Explorations in Media Visualization*

Safi Waseem, *Blind Browsing on Hand-Held Devices: Touching the Web... to Understand it Better*

Yousuf Bilal and Conlan Owen, *Constructing Narrative Visualizations as a means of Increasing Learner Engagement*

DataWiz 2014 Program Committee

Luca Maria Aiello (Yahoo! Research, Spain)

Sebastiano Battiato (University of Catania, Italy)

Federica Cena (University of Turin, Italy)

Giovanni Luca Ciampaglia (Indiana University, USA)

Emilio Ferrara (Indiana University, USA)

Bruno Gonçalves (Aix-Marseille Université, France)

Fabio Malini (Universidade Federal do Espírito Santo, Brasil)

Marco Quaggiotto (ISI Foundation, Italy)

Everardo Reyes García (Université Paris 13, France)

Rossano Schifanella (University of Turin, Italy)

Marcella Tambuscio (University of Turin, Italy)

Lilian Weng (Indiana University, USA)

A Visualization Platform For Exploring Cooperation

Remy Cazabet
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, japan
remy.cazabet@gmail.com

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, japan
takeda@nii.ac.jp

ABSTRACT

In this paper, we present a platform designed to explore visually massive cooperation between individuals. With the increasing importance of the Internet, new types of cooperation are becoming common, in which hundreds, thousands or millions of individuals act together in interaction, and produces content in a decentralized manner. As these processes are happening in real-time and without organization, individuals involved in them often do not have a clear vision of what is happening, or even which role they play in it. The visualization we propose would allow users to take back the power of understanding the processes to which they participate in. We combine time series visualization, together with custom network visualization, in a way generic enough to adapt to many situations, while offering numerous possibilities.

1. INTRODUCTION

Since the advent of the digital era, both the technical possibilities and the introduction of new behaviors have participated in the production of large databases storing tremendous amounts of varied information. Recent hot topics such as Big Data, Complex Systems and Network Analysis have been stimulated by this new access to information. One particular topic of interest is the study of how crowds are involved in massive generation of content, whether it be on Wikipedia, Twitter, Facebook, YouTube, or even through the publication of ever growing number of scientific publications. If these datasets are a stimulating opportunity, they are also a challenge. While many research has been done on these topics, we feel there is no simple, generic method to explore this decentralized creation of content, and in particular its dynamic. The platform we propose is generic enough to take input from many kinds of sources, such as scientific publications, online social networks, and many others. The platform is developed with internet based tools only, and could therefore be adapted to provide a user-friendly interface to explore a large dataset of content creation available on the internet.

1.1 Related Works

Several visualizations have been proposed to understand complex systems and large data in general. We introduce the most closely related to our proposition.

ThemeRiver [9] is probably the most famous of these. It allows to represents the dynamic of topics in large collections of documents.

History flows [16]also focuses on dynamic aspects. It is a tool to visualize cooperation and conflict between authors in the process of collaboration, in particular on the web.

In the work by Rosvall et al.[12], alluvial diagrams are used to represent the evolution of communities in networks, and is applied in particular for the visualization of the evolution of research topics in science.

On a more static perspective, numerous tools, frameworks and softwares have been proposed to represent networks in the best possible way. We can cite some of them, among the most famous ones: Gephi[2], Cytoscape[13], Tulip [1].

Several works have also been done on the visualization of dynamic networks; we can cite [3] as a reference on the domain.

The tools we have cited above are either specialized on the visualization of longitudinal aspects, but without information on the internal structure, or, on the contrary, represent this internal structure (network visualization), but only with a static point of view. Our platform is designed to encompass both aspects.

2. MASS COOPERATION DATASETS

In order to illustrate the possibilities and possible practical applications of the tools presented in this paper, we applied them to three large datasets from different fields. In this section, we will present briefly these datasets, and the type of data we extract from them.

For a dataset to be visualized using our platform, it needs to be composed of several productions, that we call Cooperative Productions (CPs). It can be a video, an article, a website, a message, or any other item which can make a reference and be referenced. These CPs are defined by the following properties:

- Name
- Time of publication
- Category (a chain of character, can be omitted)
- List of references it makes to other CPs

Additionally, we need to group these CPs in Cooperation processes. A cooperation process is a set of CPs corresponding to a same topic, a same goal, or any other way of grouping them relevant to the studied dataset. In the following sections, we will detail these properties in 3 example datasets.

2.1 NicoNico

NicoNico, or Nico Nico Douga, is a Japanese video-sharing platform, with functionalities similar to those of YouTube. With officially more than 20 Million registered users, and being ranked among the top 15 most visited websites of Japan, it is a major Web 2.0 platform. It is especially famous for the important community of people cooperating in the creation of complex Music Videos centered on the character of Hatsune Miku. Starting from an original song, many people create videos based on it, with innovation such as dancing, singing, creating new graphics, etc. More information about this character and phenomenon can be found in [8, 10, 6].

We use the dataset described in [7] which covers all 2,622,495 videos published on the network between January 2007 and December 2012.

Definition of a cooperation process
In NicoNico, tags are associated with videos. We automatically detect tags corresponding to songs with more than 500 related videos. These videos compose the cooperation processes.

Definition of a CP

- Name : Name of the Video
- Time : Upload time
- Category : extracted from keywords, examples are: Dancing, Singing, 3D, Animation...
- References: authors include references to other videos in their comments.

Statistics
We obtain 165 cooperation processes, composed by 500 to 7654 videos, with an average of 865 videos.

2.2 Twitter

Twitter is one of the most famous and largest Online Social Networks. In this paper, we consider the diffusion of a particular tweet as our cooperation processes. We used a dataset covering the period between March 5, 2011 and March 24, and which covers most tweets published in Japan during this period. Authors of this dataset claim to have validated that 80% to 90% of all published tweets appear in their dataset. For more information, please refer to [15].

Definition of a cooperation Process
We first counted for each tweet in our dataset the number of time they were retweeted, following the method described in [4]. For all tweets retweeted more than 500 times, we collect all the involved tweets and their information. Each of these sets of tweet form a cooperation flow.

Definition of a CP

- Name : Retweeter's name
- Time : Time of the Retweet
- Category : Distance in the follower network between original author and retweeter
- References: a retweet

Statistics
45 cooperation processes corresponding to retweet chains are detected, involving between 500 and 2100 tweets, with an average of 755 tweets.

2.3 DBLP

Massive cooperation predates the apparition of the World Wide Web. Thousands of researchers around the world cooperate to improve the global scientific knowledge. We use as a dataset the DBLP database [11], and in particular the version including links between papers, as described in [14]. This database is composed of 2,084,055 articles linked by 2,244,018 citations.

Definition of a cooperation Process
As we lack topic information, we define a cooperation process for each article, with all other papers making a direct reference to it composing the cooperation processes. This definition is not perfect, but, as we know that seminal papers tend to act as "flags", that must be cited by everyone working on a specific topic, looking at all papers citing a seminal one is an approximation of a group of works in the same topic. We filtered out all cooperation processes with less than 500 elements.

Definition of a CP

- Name : Publication Title
- Time : Date of Publication
- Category : Venue of publication
- References: a citation to another paper

Statistics
After filtering, we obtained 41 citation flows, composed of between 500 and 3651 papers, with an average of 664 papers.

3. DESCRIPTION OF THE PLATFORM

The platform we propose is composed of two parts: the time series visualization and the cooperation flow visualization. The time series provides a global understanding of the different cooperation processes studied, together with global

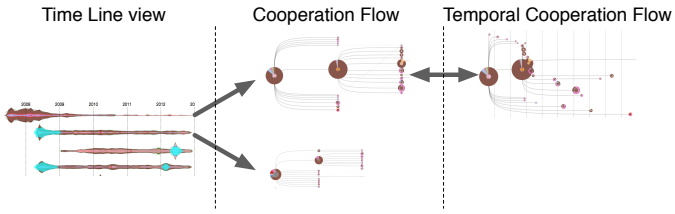


Figure 1: Schema of the possible navigation between the displays of Cooperation Explorer

indicators on them. In this view, only the global properties are represented, not the individual agents and their interactions. From this global view, it is then possible to select any cooperation process and to visualize its inner details in the cooperation flow view.

In this second tool, in which the details of the cooperation is displayed, several options are possible such as positioning according to time or to step of cooperation, selecting the number of nodes displayed, etc. The navigation between these different displays in represented in Fig. 1

3.1 Temporal trends

When we are interested in a cooperation process, it is often useful to have first a global vision of it. We would like to be able to answer general questions such as: when did this process started? Is it already finished? Is it becoming more or less popular? Are there some patterns in its popularity? These are the global properties of this particular cooperation.

3.1.1 Macro-level visualization: time series

The visualization we propose excludes the role of each specific element, to represents the process as a whole. To do so, we choose to transform our data in time series, as much work exists on the topic of time series analysis. For a given dataset, we define a time step, which can be any period of time (minute, day, year, etc.) and count the number of CPs published for each category in each time step. For each Cooperation Flow, we obtain as many time series as there are categories. We display them as a shape, as shown in Fig. 7. The shape is constructed as a cumulative area chart augmented with a mirror image of itself, to have a symmetric shape. The lecture of it is identical to a normal cumulative area chart. We choose this shape instead of a normal cumulative area chart because we want to represent several of these shapes on a same plot with a single time axis. Therefore, the shape is not framed by the axis, and when displayed on top of each other, it becomes more natural to have a horizontally symmetrical shape, as represented on Fig. 8. A similar observation has been done by the authors of ThemeRiver [9].

By displaying several shapes on the same chart, we are able to visually compare them. Examples of interesting observable facts include (but are not limited to):

- The relative importance of different categories along time
- The presence of bursts at a particular location, or following a fix period
- Differences between cooperation processes starting at different times

We complete this tool with some metrics:

3.1.2 metrics and graphics

Lifespan

For each cooperative Process, we compute its lifespan, defined as the time between the first not null value of the time series to the last occurrence of 3 consecutive not null values. This limit is arbitrary, but the objective is to give an end to a time series, potentially infinite, as a new CPs can always occurs in the future. If these 3 non-null values are the last 3 values of the time series, we consider the cooperation process as "still alive". The distribution of the lifespans is displayed as a bar chart.

Normalized centroid

We compute the normalized centroid of each cooperative flow. The centroid of the time series is the step such as there is as many CPs before and after it. We normalize it by computing:

$$NormalizedCentroid = \frac{centroidTime - birthTime}{deathTime - birthTime}.$$

A normalized centroid inferior or superior to 0.5 reflect the fact that most of the CPs where produced in the beginning or in the end of the lifespan of the cooperation process. The distribution of the normalized centroid is displayed as a bar chart.

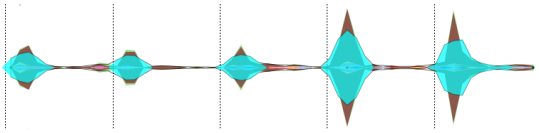


Figure 2: Visualization of periodic bursts in a temporal trend. Detected bursts appear in translucent blue color. In this case, we can observe yearly events.

Burst detection

Burst detection is a common problem on time series. A burst is defined as a period of time during which the time series reach temporarily exceptionally high values. In a cooperation flow, such a burst can typically appears in the beginning (initial burst), at a given moment, driven by internal events (new popular CP), or external factors. An interesting case is when this external factor is not unique but periodic, typically daily or yearly events. We therefore implemented a research of such periodic bursts. We implemented the burst detection with a simple but effective technique, presented in [17]. We represent the bursting period with a translucent color as seen in 2. We compute normalize burst positions in a similar manner as we computed normalized centroid, and the summary of the most common burst positions detected is also represented as a bar chart.

We found 5 cooperation processes with periodic bursts in the NicoNico dataset, and we checked that all of them cor-

responded to yearly events (songs about Christmas, Halloween, etc.).

3.2 Micro-level visualization

Whereas the time series visualization allow us to have a quick understanding of global properties, it is often useful to have more insights in the details of what is happening inside each cooperative topic. In this second display, we combine a visualization called cooperation flow together with some alternatives displays and indicators, each of them emphasizing one aspect of the studied cooperative topic.

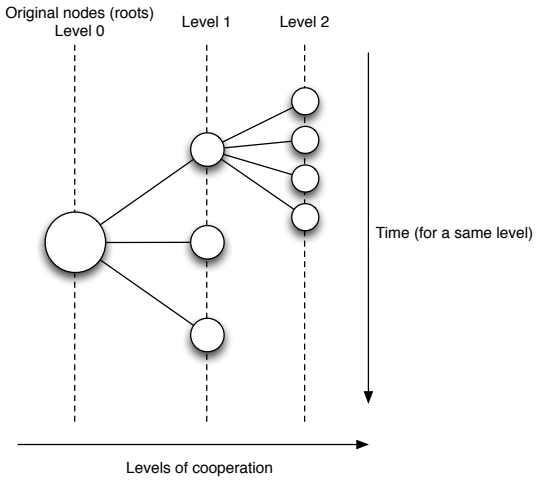


Figure 3: Mechanism of the cooperation flow visualization.

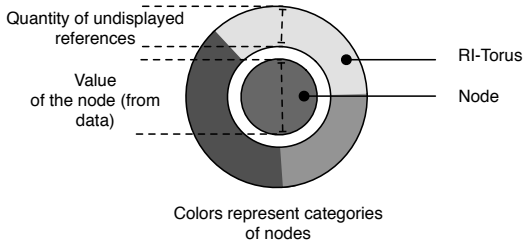


Figure 4: Schema of the representation of a node

3.2.1 *Cooperation Flow*

To represent the details of the process of cooperation, we use a type of visualization described in [5]. This visualization, called Cooperation flow, allows us to represent in a single visualization the key points of the details of the process. Its mechanism is represented in Fig.3. The idea is that, through the interface, we specify the maximum number of nodes that we want to display, n. An algorithm compute which are the n most important elements for the cooperation in the current process. These nodes are then displayed, together with their relations, as a network organized by steps of cooperation. More formally, the step of a node is defined as the length of the shortest path between this node and a root, that is to say a node without any reference to other nodes. The nodes which are not considered important enough to be displayed are, however, not simply omitted. By using a feature called Reuse Indicator Torus (RI-Torus), a summary of these nodes appears around their last displayed ancestor, as summarized in Fig. 4.

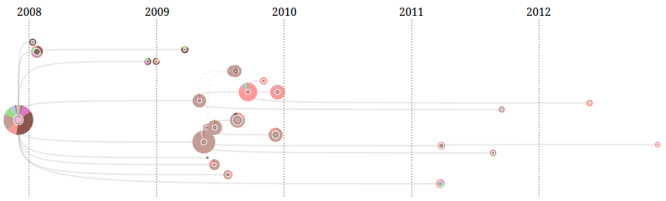


Figure 5: Example of a cooperation flow where the x position represents time

3.2.2 *Temporal Cooperation Flow*

One interesting property of this visualization is that nodes situated on overlapping y values are necessarily ordered in a chronological order from left to right. Therefore, it is possible to switch to a temporal representation without changing the y position of nodes. This is illustrated on figure 5.

3.2.3 *Complementary visualizations and metrics*

We added to this visualization a set of informative visualization and metrics, each of them focusing on a specific aspect of the cooperation. These tools are based on the same data as the cooperation flow visualization. All of these tools are not affected by the selection of nodes we make for the flow visualization, they are based on all available information.

Impact of main CPs

We observed that one characteristic which can vary greatly between cooperation flows is the importance taken by the most important productions. In some cases, a single production, or a small subset of them, can generate most of the CPs, that is, most of the CPs will directly reference it as a unique source, either during the whole lifetime of the flow, or just during a given period. To study this, we propose a visualization in stacked area of the impact along time of the top 5 nodes, toped by the impact of all remaining nodes (Fig. 6). The lifetime of the flow is split in 10 sections. The impact of a given CP during a given section is computed as the number of CPs published during this period that reference it. We use a black and white scale to avoid confusion with the categories of CPs, already represented by colors.

Together with this visualization, we propose a metric to measure this effect, called CSC, for Cooperation Source Concentration.

$$CSC = \frac{\sum_{v \in Top1} |\{u : (u,v) \in E\}|}{|V| - 1}$$

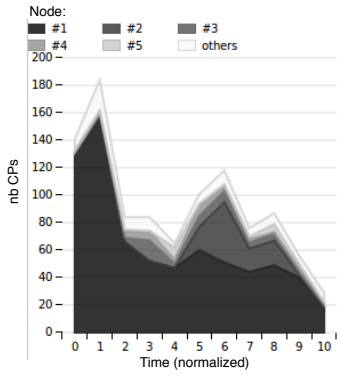
Where $Top1$ is the set of the 1% nodes with the highest in-degree. This metric vary between lim_0 and 1, where lim_0 is the case where all nodes have the same in-degree, and 1 is reached when all nodes are successors of a single original source. (star-like network) We give the average values of CSC for our 3 datasets in Table 1. We can observe large differences, with NicoNico having the strongest CSC and Twitter the lowest.

Sustainability of the cooperation

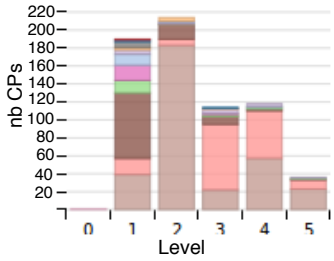
We observed that in some cooperation flow, there is not much cooperation after the first few levels -the number of CP by level follows a fast shrinking trend- while, in others, it is not the case. This might reflect the ability to renew the

	NicoNico	Twitter	DBLP
Average CSC	0.92	0.21	0.65

Table 1: Average value of CSC by dataset. CSC represents how important is the role of the top 1% users in the cooperation.



(a) Impact of the main CPs along time, in a particular cooperation flow. In this example, we can see that 2 or 3 nodes are the source of most cooperation. For example, the increasing number of videos in period 6 is mainly due to the popularity of a single node.



(b) Sustainability of the cooperation. In this case, we can observe a shift in the categories of CPs as with progress in the levels of cooperation. This pattern is common on NicoNico.

Figure 6: Additional analysis tools

interest in the trend by new CPs. We propose a visualization of this effect by a stacked bar chart graph (Fig. 6). Each bar represent a level, and we simply count the number of videos of each type published in each level. Together with the general trend, this chart allows to see a change in the categories correlated with the level. The color used for the categories are coherent with the ones used in the cooperation flow chart.

The indicator we propose to summarize this chart is SC, Sustainability of Cooperation. It is defined as the average of the variations of the number of CPs between successive levels, pondered by the number of CPs in the first of the two:

$$SC = \frac{\sum_{i=1}^{nl-1} \frac{nbCP(i+1)}{nbCP(i)} * (nbCP(i+1) + nbCP(i))}{nbCP(1) + 2\sum_{i=2}^{nl-2} nbCP(i) + nbCP(nl-1)}$$

with $nbCP(i)$ the number of CPs at level i, and nl the number of levels. SC=0 if there is no production after the first level. SC >1 if the number of CPs tends to grow with each level. The lower the SC value, the less CPs tend to generate new cooperation. In Table 2, we represent the average value of SC for our datasets.

	NicoNico	Twitter	DBLP
Average SC	0.23	0.39	0.44

Table 2: Average value of SC by dataset. SC represents the average ratio between the number of videos published at step i and $i+1$.

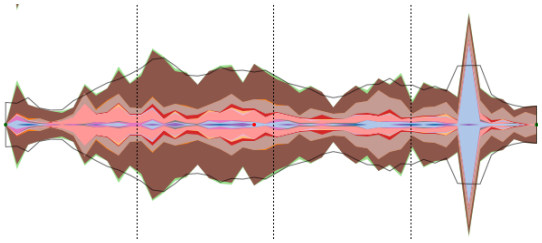


Figure 7: Visualization of the temporal trend of a cooperation flow

4. EXAMPLE VISUALIZATIONS

In this section, we briefly present some examples to show the interest of our visualization.

4.1 Temporal trends

Our visualization allows to display on a same timeline the time series of several temporal trends. We can therefore compare them, observe some typical behaviors or spot outliers. In Fig. 8, we show an example of this view on our example dataset. We can observe very different properties. For instance, in Twitter, we see a typical bursting behavior, followed by a rapid decay. Most of the productions, i.e., retweets, occur in the beginning. On Nico Nico, the trends are more long lasting, bursts are not as important. People continue to publish videos at the same rate for years. Finally, in the citation dataset, we observed more varied patterns, and even some "increasing" trends, for which the number of papers published increase from years to years.

4.2 Cooperation flows

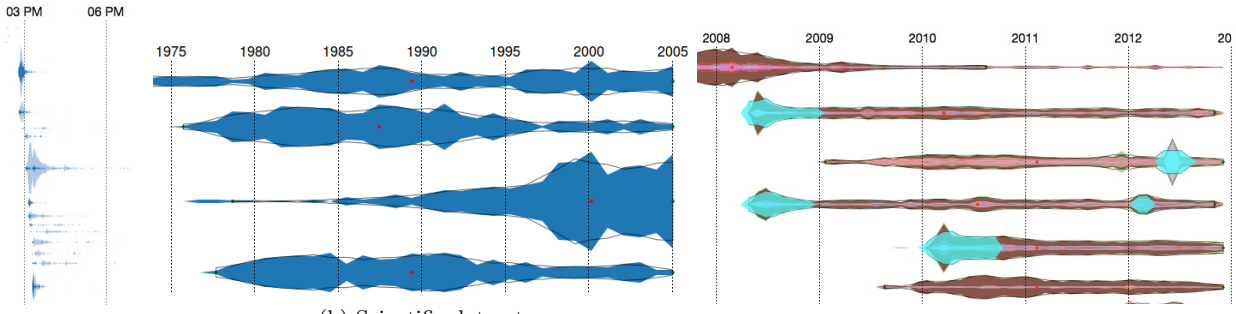
4.2.1 Deep study of one dataset: NicoNico

NicoNico is the richest and the most complex of our datasets. In fig. 9, we show 2 typical flow from this network. We can make the following observations, also valid on most other flows:

1. There is only one original source, and most of the cooperation is made directly from this source, as we can judge by the large RI-Torus
2. Most important nodes for the collaboration are on the first level, they directly reference the original node only
3. The cooperation is more wide than deep, there is not much cooperation at a level greater than 3.
4. Although many categories (colors) are present, each node seems to generate a specialized cooperation: RI-Torus are mostly of a single color, not always the same.
5. There is no strong correlation between the number of view of a video (area of inner circle) and its capacity to generate cooperative behavior (torus area)

4.2.2 Comparison of datasets

In fig. 10, we present two visualizations typical of the other datasets. We can immediately spot some differences. In the tweet dataset, cooperation is deeper, and we tend to see the formation of chains, long but without many bifurcations. More important nodes are not necessarily situated



(a) Twitter dataset

(b) Scientific dataset

(c) Nico Nico Dataset

Figure 8: Examples of temporal trends

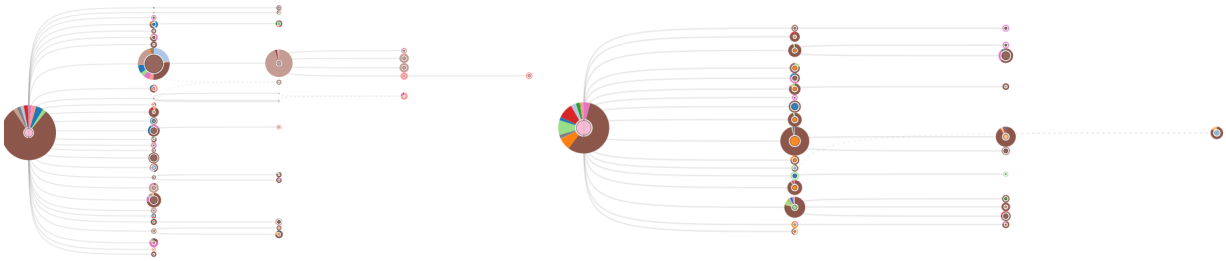
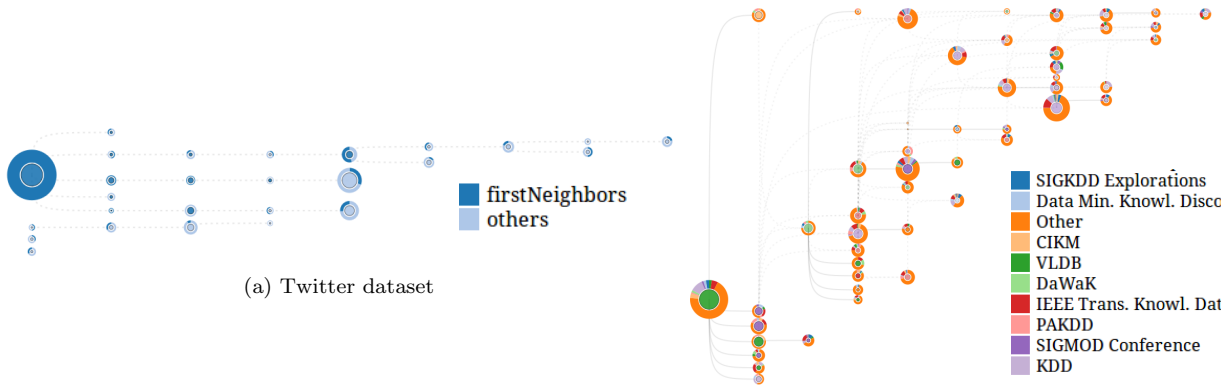


Figure 9: Examples of typical cooperation flow in NicoNico



(a) Twitter dataset

(b) DBLP dataset

Figure 10: Examples of typical cooperation flow in Twitter and DBLP

at the first step, but can occur deeper. There seems to be a stronger relation between the popularity of the node and its role in the cooperation. There is not a single source.

In the citation dataset, we immediately spot a large number of nodes making references to several others. These nodes with many references are important in the cooperation. Nodes at a deep level seem to generate as much cooperation as those in the first levels. There also seems to be a lesser concentration in the cooperation generation: a larger fraction of nodes are referenced by other important nodes, and the gap is less important between the top influential nodes and the ordinary ones. Exploring in further details the properties of the different datasets is beyond the scope of this paper.

5. CONCLUSION

In this paper, we have presented a platform to explore mass cooperation, and a set of tools to explore different aspects of this type of cooperation. Our conception of such visualization was driven by our previous experiences in the exploration of large datasets formed by cooperation, and the difficulties encountered to understand the underlying mechanisms.

We also presented some complementary visualizations and metrics that focus on several aspects of the data, with different granularities, and can also help to apprehend it.

In the future, we hope that other researchers will use this platform and help to improve it, either by their remarks or extending the possibilities. In this prospect, we release its source code, altogether with an interactive online version, so as interested researchers could work with it as easily as possible. In particular, it could be interesting to add metrics and statistics, such as a one could choose the more interesting indicators in his case. The source code and browsable example is available on the website of the first author.

Another future possibility is to propose Internet applications based on this visualization to the destination of final end users. For example, one can think of a plug-in for Google Scholar allowing to browse research topics.

6. ACKNOWLEDGMENTS

We thank Fujio Toriumi for collecting the Twitter dataset, and allowing us to make use of it in this work.

7. REFERENCES

[1] D. Auber. Tulip, a huge graph visualization framework. In *Graph Drawing Software*, pages 105–126. Springer, 2004.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *ICWSM*, pages 361–362, 2009.

[3] S. Bender-deMoll and D. A. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7(2):1–38, 2006.

[4] R. Cazabet, N. Pervin, F. Toriumi, and H. Takeda. Information diffusion on twitter: everyone has its chance, but all chances are not equal. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*, pages 483–490. IEEE, 2013.

[5] R. Cazabet and H. Takeda. Understanding mass cooperation through visualization. *ACM Conference on Hypertex and Social Media*, 2014.

[6] R. Cazabet, H. Takeda, M. Hamasaki, and F. Amblard. Using dynamic community detection to identify trends in user-generated content. *Social Network Analysis and Mining*, 2(4):361–371, 2012.

[7] M. Hamasaki and M. Goto. Songrium: a music browsing assistance service based on visualization of massive open collaboration within music content creation community. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 4. ACM, 2013.

[8] M. Hamasaki, H. Takeda, and T. Nishimura. Network analysis of massively collaborative creation of multimedia contents: case study of hatsune miku videos on nico nico douga. In *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, pages 165–168. ACM, 2008.

[9] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[10] H. Kenmochi. Vocaloid and hatsune miku phenomenon in japan. *Proc. of InterSinging 2010*, pages 1–4, 2010.

[11] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer, 2002.

[12] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[13] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[15] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1025–1028. International World Wide Web Conferences Steering Committee, 2013.

[16] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM, 2004.

[17] Y. Zhu and D. Shasha. Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 336–345. ACM, 2003.

A Data Extraction and Visualization Framework for Information Retrieval Systems

Alessandro Celestini
Institute for Applied Computing, National Research Council of Italy
a.celestini@iac.cnr.it

Antonio Di Marco
Institute for Applied Computing, National Research Council of Italy
a.dimarco@iac.cnr.it

Giuseppe Totaro
Department of Computer Science, University of Rome "Sapienza"
totaro@di.uniroma1.it

ABSTRACT

In recent years we are witnessing a continuous growth in the amount of data that both public and private organizations collect and profit by. Search engines are the most common tools used to retrieve information, and more recently, clustering techniques showed to be an effective tool in helping users to skim query results. The majority of the systems proposed to manage information, provide textual interfaces to explore search results that are not specifically designed to provide an interactive experience to the users.

Trying to find a solution to this problem, we focus on how to extract conveniently data from sources of interest, and how to enhance their analysis and consultation through visualization techniques. In this work we present a customizable framework able to acquire, search and interactively visualize data. This framework is built upon a modular architectural schema and its effectiveness will be illustrated by a prototype implemented for a specific application domain.

Keywords

Data Visualization, Data Extraction, Acquisition.

1. INTRODUCTION

The size of data collected by private and public organizations is steadily growing and search engines are the most common tools used to quickly browse them. Many works, in different research areas, face the problem of how to manipulate such data and to transform them into valuable information, by making them *navigable* and easily searchable. Clustering techniques have been shown to be quite effective to that purpose and have been thoroughly investigated in the past years [17, 18, 2]. However the majority of currently available solutions (e.g., Carrot2[1], Yippy[2]) just supply textual interfaces to explore search results.

In recent years, several works studied how users interact with interfaces during exploratory search sessions, reporting useful results about their behavior [12, 11]. These works show that users spend the majority of their time looking at the results and at the facets, whereas only a neglectable amount of time for looking at the query itself [11] underlining the importance of user interfaces development. According to those works, it is clear that textual interfaces are not very effective to improve exploratory search, so a different solution has to be applied.

Data visualization techniques seem to be well suited to pursue such goals. Indeed, visualization offers an easy-to-use, efficient, and effective method capable to present data to a large and diverse audience including users without any programming background. The main goal of such techniques is to present data in a fashion that supports intuitive interaction to spot patterns and trends, thus making the data usable and informative. In this work we focus on data extraction and data visualization for information retrieval systems, i.e., how to extract data from the sources of interest in a convenient way, and how to enhance their analysis and consultation through visualization techniques. To meet these goals we propose a general framework, presenting its architectural schema composed of four logic units: acquisition, elaboration, storage, visualization. We also present a prototype developed for a case study. The prototype has been implemented for a specific application domain and is available online.

The rest of the paper is organized as follows. Section 2 discusses some frameworks and platforms related to our study. Section 3 presents the framework architectural schema. Section 4 describes a prototype through a case study, and finally, Section 5 concludes the paper suggesting directions for future works.

2. RELATED WORK

In this section we discuss some works proposing frameworks and platforms for data visualization.

WEKA [9] is a Java library that provides a collection of state-of-the-art machine learning algorithms and data processing tools for data mining tasks. It comes with several graphical user interfaces, but can also be extended by using a simple API. The WEKA workbench includes a set of visualization tools and algorithms for classification, regression, attribute selection, and clustering, useful to discover and understand data.

Orange [6] is a collection of C++ routines providing a set of data mining and machine learning procedures which can be easily combined in order to develop new algorithms.

[1] http://project.carrot2.org
[2] http://www.yippy.com/

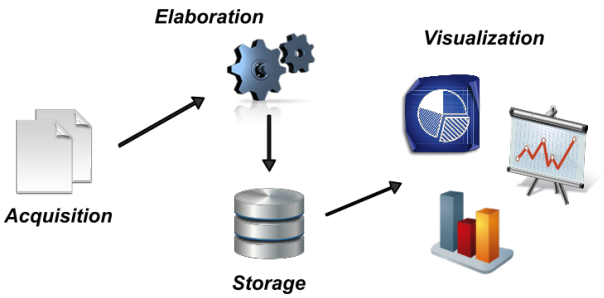


Figure 1: Architectural Schema

The framework allows to perform different tasks including data input and manipulation, methods for developing classification models, visualization of processed data, etc. Orange provides also a scriptable environment, based on Python, and a visual programming environment, based on a set of graphical widgets.

While WEKA and Orange contain several tools to deal with data mining tasks, our aim is to improve information retrieval systems and user data understanding through visualization techniques. Basic statistical analysis on data, should be implemented by charts through interactions patterns, so that could be performed directly by users.

In [8] authors present FuseViz, a framework for Web-based fusion and visualization of data. The framework provides two basic features: fusion and visualization. FuseViz collects data from multiple sources and fuses them into a single data stream. The joint data streams are then visualized trough charts and maps in a Web page. FuseViz has been designed to operate in a smart environment, where several deployed probes sense the environment in real time, and the data to visualize are live time series.

The Biketastic platform [16] is an application developed to facilitate knowledge exchange among bikers. The platform enables users to share routes and experience. For each route Biketastic captures location, sensed data and media. Such information are recorded while participants ride. Routes' data are then managed by a backend platform that makes visualizing and sharing routes' information easy and convenient.

FuseViz and Biketastic share the peculiarity of being explicitly designed to cope with a specific task in a particular environment. The proposed schemas could be re-implemented in different applications, but there is not a clear extension and adaptation procedure defined (and possibly supported) by the authors. Our aim is to present a framework that: *a*) can be easily integrated with an existing information retrieval system *b*) provides a set of tools to profitably extract data from heterogeneous sources *c*) requires minimum effort to produce new interactive visualizations.

3. FRAMEWORK OVERVIEW

Our framework adheres to a simple and well-known schema (shown in Figure 1) structured in four logic units:

1. **Acquisition**: aims at obtaining data from sources;
2. **Elaboration**: responsible for processing the acquired data to fit operational needs;
3. **Storage**: stores the data previously processed in persistent way and make them available to the users;
4. **Visualization**: provides a visual representation of data.

Actually the framework is mainly focused on the acquisition and visualization stages, whereas the other ones are reported as part of the architecture but are not implemented by us. From an engineering perspective, both middle stages (elaboration and storage) are considered as black-box components: only their input and output specifications must be available. All logic units play a crucial role for visualizing data thus we describe them according to the purposes of our framework.

3.1 Acquisition

This component is in charge of collecting and preprocessing data. Given a collection of documents, possibly in different formats, the acquisition stage prepares data and organizes them to feed the elaboration unit.

Data acquisition can be considered the first (mandatory) phase for any data processing activity that anticipates the data visualization. Cleveland [5] and Fry [7] examine in depth the logical structure of visualizing data by identifying seven stages: *acquire*, *parse*, *filter*, *mine*, *represent*, *refine*, and *interact*. Each stage in turn requires to apply techniques and methods from different fields of computer science.

The seven stages are important in order to reconcile all scientific fields involved in data visualization especially from the logical point of view. However, regarding to our prototype we refer to data acquisition as a software component which is able to collect, parse and extract data in an efficient and secure way. The output of data acquisition will be a selection of well-formed contents that are intelligible for the elaboration unit.

We can collect data[3] by connecting the acquisition unit to data source (e.g., files from a disk or data over a network). The approach to data collection depends on goals and desired results. For instance, forensic data collection requires the application of scientifically sound and proven methods[4] to produce a bit-stream copy from data, that is an exact bit-by-bit copy of the original media certified by a message digest and/or a secure hash algorithm. Thus, data collection in many circumstances has to address specific issues about prevention, detection and correction of errors.

The acquired data must be parsed according to their digital structure in order to extract data of interest and prepare them for an elaboration unit. Parsing is potentially a time-consuming process especially while working with heterogeneous data formats. The parsing stage is necessary also to extract the metadata related to examined data. Both textual contents and metadata are usually extracted and stored in specific data interchange formats like JSON or XML.

Moreover, security and efficiency aspects have to be considered during the design of a data acquisition unit. However,

[3]We assume to work with static data. Static/persistent data are not modified during data acquisition, while dynamic data refer to information that is asynchronously updated.

[4]`http://dfrws.org/2001/dfrws-rm-final.pdf`

it is beyond the scope of the present work to discuss security and efficiency related issues regardless their important implications for data acquisition.

3.2 Elaboration and Storage

The elaboration unit takes as input the data extracted during the acquisition phase, so it has to analyze and extrapolate information from them. Data analysis for instance, may be performed by a semantic engine or a traditional search engine. In the former case we will obtain, as output, the documents collection enriched with semantic information, in the second case the output will be an index. Moreover, along with the analysis results, the elaboration unit may return analysis of the metadata, related to the documents, which are received as an input.

The main task of the storage unit is to store analysis results produced by the elaboration unit and make them available for the visualization unit. At this stage the main issue is to optimize data access, specifically the querying time, in order to reduce the time spent by the visualization unit retrieving the information to display. Several storage solutions can be implemented, in particular one may choose among different types of data bases [3, 13]. The traditional choice could be a relational database, but there are several alternatives, e.g., XML databases or graph databases.

3.3 Visualization

The visualization unit is in charge of making data available and valuable for the user. As a matter of fact, visualization is fundamental to transform analysis results into valuable information for the user and help her/him to explore data. In particular, the visualization of the results may help the user to extract new information from data and to decide future queries. As previously discussed, the time spent by the user looking at the query itself is negligible, whereas the time spent looking at the results and how they are displayed is long-lasting. Thus, the interface design is crucial for the effectiveness of this unit, and the guidelines outlined in [12] may became a useful guide for the design and implementation of this unit. Given the tight interaction with the user, it is quite important to take into account the response time and usability of the interface. The visualizations provided should be interactive, to enable the user performing analysis operations on data. The same data should be displayed in several layouts to highlight their different aspects. Finally, it is quite important to provide multiple filters for each visualization, in order to offer to the user the chance of a dynamic interaction with the results.

3.3.1 The "Wow-Effect"

A really-effective data visualization technique has to be developed keeping in mind two fundamental guidelines that are abstraction and correlation.

However, scientists often focus on the creation of trendy – but not always useful – visualizations that should arouse astonishment in the users who observe them, causing what McQuillan [14] defines as the *Wow-Effect*. Unfortunately, the Wow-Effect vanishes quickly and results in having stunning visualizations that are worthless for the audience. This effect is also related to the intrinsic complexity of the data generated from acquisition to visualization stage. As shown in Figure 2, the impact of original data into the total amount

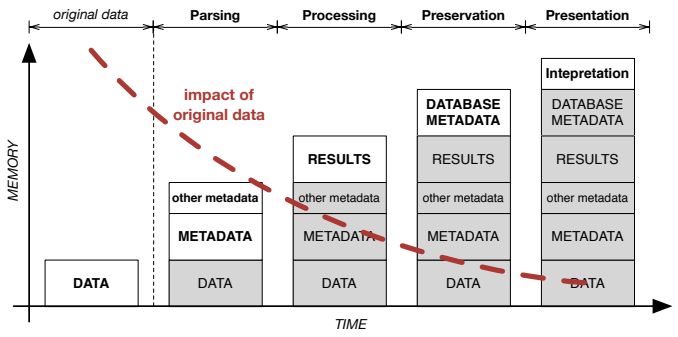


Figure 2: Data enrichment over time

of information decreases over time. Thus, we invested in effort to develop a framework able to overcome the "negative" wow effect by providing visualizations easy to use and effective.

4. CASE STUDY: 4P'S PIPELINE

In this section we present an application of the framework developed for a case study. According to the main task accomplished by each framework unit, we named the whole procedure the *4P's pipeline*: parsing, processing, preservation, and presentation.

The prototype is a browser based application available online[5]. The data set used for testing the 4P's pipeline is a collection of documents in different file formats (e.g., PDF, HTML, MS Office types, etc). The data set was obtained by collecting documents from several sources, mainly related to news in English language.

4.1 Parsing task

The acquisition unit is designed to effectively address the issues discussed in Section 3.1. Parsing is the core task of our acquisition unit and for its implementation we exploited the Apache Tika[6] framework. The Apache Tika is a Java library that carries out detection of document type and the extraction of both metadata and structured textual content. It uses existing parser libraries and supports most data formats.

4.1.1 Tika parsing

Tika is currently the de-facto "babel fish", performing automatic text extraction and content analysis of more than 1200 data formats. Furthermore there are several projects that aim at expanding Tika to handle other data formats.

Document type detection is based on a taxonomy provided by the IANA media types registry[7]that contains hundreds of officially registered types. There are also many unofficial media types that require attention, so Tika has its own media types registry that contains both official registered types and other, widely used albeit unofficial, types. This registry maintains information associated to each supported type. Tika implements six methods for type detection [4] respectively based on the following criteria: filename patterns, `Content-Type` hints, magic byte prefixes, character encodings, structure/schema detection, combined approaches.

[5] `http://kelvin.iac.rm.cnr.it/interface/`
[6] `http://tika.apache.org/`
[7] `http://tools.ietf.org/html/rfc6838`

The `Parser` interface is the key concept of Apache Tika. It provides a high level of abstraction hiding the complexity of different file formats and parsing libraries. Moreover, it represents an extension point to add new parser Java classes to Apache Tika, that must implement the Parser interface. The selection of the parser implementation to be used for parsing a given document may be either explicit or automatic (based on detection heuristics).
Each Tika parser allows to perform text (only for text-oriented types) and metadata extraction from digital documents. Parsed metadata are written to the `Metadata` object after the `parse()` method returns.

4.1.2 *Acquisition unit in detail*

Our acquisition unit uses Tika to automatically perform type detection and parsing, against files collected from data sources, by using all available detectors and parser implementations. Although Tika is, to the best of our knowledge, the most complete and effective way to extract text and metadata from documents, there are some situations where it could not accomplish its job, for example when Tika fails to detect the document format or, even if it correctly recognizes the filetype, when an exception occurs during parsing. The acquisition unit handles both situations by using alternative parsers which are designed to work with specific types of data (see figure 3):

- Whenever Tika is not able to detect a file because either it is not a supported filetype or the document is not correctly detectable (for example, it has a malformed/misleading `Content-Type` attribute), the examined file is marked as `application/octet-stream`, i.e., a type used to indicate that a body contains arbitrary binary data. Therefore, the acquisition unit processes documents whose the exact type is undetectable by using a customized set of ad-hoc parsers, each one specialized to handle specific types. For instance, Tika does not currently support Outlook PST files, so they are marked as `octet-stream` subtypes. Then, the acquisition unit analyzes the undetected file by using criteria as extension pattern or more sophisticated heuristics and finally it sends the binary data to an ad-hoc parser based on the *java-libpst*[8] library.
- During parsing, even though a document is correctly detected by Tika, some errors/exceptions can occur, interrupting the extraction process related to the target file. In this case, the acquisition unit tries to restart the parsing against the file that has caused a Tika exception by using, if available, a suitable parser selected from an ad-hoc parsers list.

The acquisition unit extracts metadata from documents according to a unified schema based on basic metadata properties contained in the `TikaCoreProperties` interface, which all (Tika and ad-hoc) parsers will attempt to extract. A unified schema is necessary in order to have a unique experience with searching against metadata properties. A complete and more complex way to address "metadata interoperability" consists in applying schema matching techniques in order to provide suitable metadata crosswalks.

[8] `https://code.google.com/p/java-libpst/`

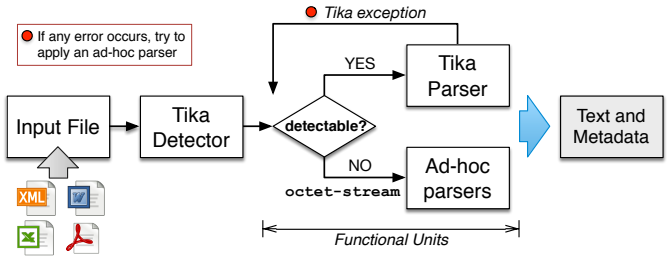


Figure 3: Acquisition unit

4.2 Processing and Preservation tasks

The second and the third tasks are respectively the processing and the preservation of data. The elaboration and storage units which perform these tasks are tightly coupled. All processed data must be stored in order to preserve the elaboration results in a persistent way. They work by using a simple strategy like *Write-Once-Read-Many* pattern, where the visualization unit plays the reader role.

4.2.1 *Elaboration unit*

The elaboration unit is formed by the semantic engine Cogito[9]. Cogito analyzes text documents, and is able to find hidden relationships, trends and events, transforming unstructured information into structured data. Among the several analysis it identifies three different types of entities (people, places and companies/organizations), categorizes documents on the basis of several taxonomies and extract entities co-occurrences. Notice that this unit is outside the framework despite we included it in the architectural schema. Indeed, we do not take care of the elaboration unit design and development, we consider it as given. This unit is the entity with which the framework interacts and to which the framework provides functionalities, i.e., text extraction and visualization.

4.2.2 *Storage unit*

As storage unit we resorted to BaseX[10], an XML data base. BaseX is an open source solution released under the terms of the BSD License. We decided to use an XML data base because the results of the elaboration unit are returned in XML format. Moreover, the use of an XML data base helps to reduce the time for XML documents manipulation and processing, compared to a middleware application [10, 15]. An XML data base has also the advantage of not constraining data to a rigid schema, namely in the same data base we can add XML documents with different structures. Thus, the structure of the elaboration results can change without effecting the data base structure itself.

4.3 Presentation task

For the development of the visualization unit we used D3.js[11] [1], a JavaScript library. The library provides several graphical primitives to implement visualizations and uses only web standards, namely HTML, SVG and CSS. With D3 it is possible to realize multi-stage animations and interactive visualizations of complex structures.

[9] `http://www.expertsystem.net`
[10] `http://basex.org`
[11] `http://d3js.org`

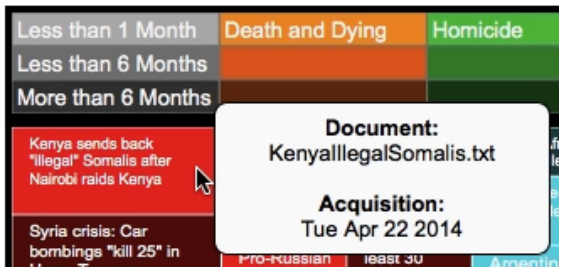


Figure 4: Treemap with category zooming

Figure 5: Geographic visualization and country selection

To improve data retrieval, we realized several visualization alternatives that exploit Cogito's analysis results. Figure 4 shows a treemap visualization that displays a documents categorization, notice that the same document may fall in different categories. Not all categories are displayed, only eight among the most common ones. The categories reported are selected on the basis of the number of documents contained in the category itself. The treemap visualization is quite effective in providing a global view of the data set. Our implementation enables also a category zooming to restrict the set of interest, i.e., clicking on a document the visualization displays only the documents in the same category. Moreover, the user is able to retrieve several information such as the document's name, part of the document content and the document's acquisition date, directly from the visualization interface. Figure 5 shows a geographic visualization that displays a geo-categorization of documents. The countries appearing in the documents are rendered with a different color (green), to highlight the difference respect to the others. The user can select each green country to get several

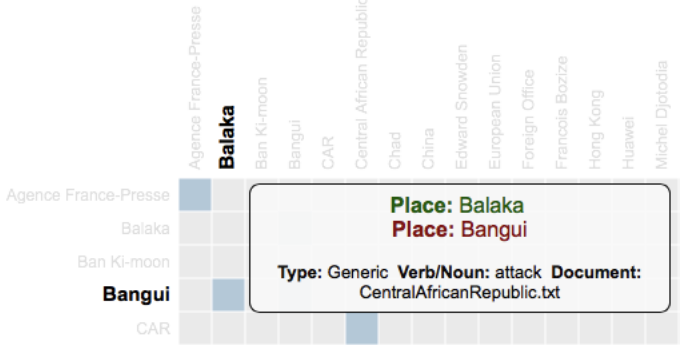


Figure 6: Co-occurrences matrix.

information that are reported inside a tooltip as shown in figure. For each country are reported general information such as capital's name, spoken languages, population figures, etc. Such information do not come from the Cogito analysis, but are added to enrich and enhance the retrieval process carried out by users. The tooltip reports also the list of documents in which the country appears and the features detected by Cogito. Features are identified according to a specific taxonomy and for each country are reported all the features detected inside the documents related to that country. Moreover, this visualization displays geographic locations belonging to the country, possibly identified during the analysis, e.g. rivers, cities, mountains, ecc. Figure 6 shows the visualization of entities co-occurrence (only a section of the matrix is reported in figure). Three types of entities are identified by Cogito, that are places, people, organizations. All entities are listed both on rows and columns, when two entities appear inside the same document the square at the intersection is highlighted. The color of the squares is always the same, but the opacity of each square is computed on the basis of the number of co-occurrences. Thus, the higher the number of co-occurrences, the darker the square at the intersection. Furthermore, a tooltip for each highlighted square reports the type of the two entities, information about the co-occurrence and the list of documents in which they appear. Specifically, the tooltip reports the verb or noun connecting the entities and some information about the verb or noun used.

Figure 7 shows a force directed graph that displays the relations detected among the entities identified in the documents. Each entity is represented by a symbol denoting the entity's type. An edge connects two entities if a relation has been detected between them, self-loop are possible. Edges are rendered with different colors based on relations' type. The legend concerning edges and nodes is reported on top of the visualization. A tooltip reports some information about the relations. In particular, for each edge is reported the sentence connecting the entities, the verb or noun used in the sentence and the document's name in which the sentence appear. Instead for each node a tooltip reports the list of document in which the entity appears. Furthermore, for each visualization, the user may apply several filters. In particular, we give the possibility to filter data by acquisition date, geographic location, nodes' types (co-occurrence matrix and force directed graph), relations' type (force directed graph), categories (treemap).

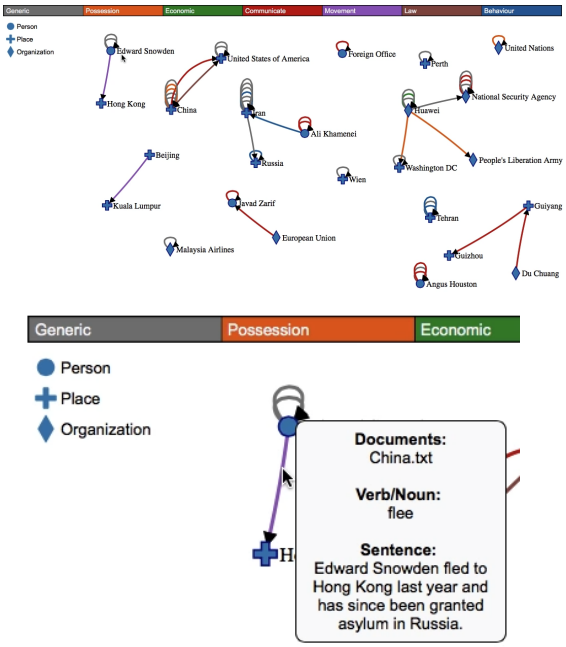


Figure 7: Entity-relations force directed graph

5. CONCLUSIONS

The interest in data visualization techniques is increasing, indeed these techniques are showing to be a useful tool in the processes of data analysis and understanding. In this paper we have discussed a general framework for data extraction and visualization, whose aim is to provide a methodology to conveniently extract data and facilitate the creation of effective visualizations. In particular, we described the framework's architecture, illustrating its components and its functionalities, and a prototype. The prototype represents an example of how our framework can be applied when dealing with real information retrieval systems. Moreover, the online application demo provides several visualization examples that can be reused in different contexts and application domains.

Currently we're experimenting our prototype for digital forensics and investigation purposes, aiming at providing to law enforcement agencies a tool for correlating and visualizing off-line forensic data, that can be used by an investigator even if she/he does not have advanced skills in computer forensics. As a future activity we plan to release a full version of our prototype. At the moment the elaboration engine is a proprietary solution that we cannot make publicly available, hence we aim at replacing this unit with an open solution. Finally, we want to enhance our framework in order to facilitate the integration of data extraction and data visualization endpoints with arbitrary retrieval systems.

Acknowledgements

We would like to express our appreciation to Expert Systems for support in using Cogito. Moreover, financial support from EU projects HOME/2012/ISEC/AG/INT/4000003856 and HOME/2012/ISEC/AG/4000004362 is kindly acknowledged.

6. REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 Data-Driven Documents. *IEEE TVCG*, 17(12):2301–2309, Dec 2011.

[2] C. Carpineto, S. Osiński, G. Romano, and D. Weiss. A survey of web clustering engines. *ACM Comput. Surv.*, 41(3):1–17, Jul 2009.

[3] R. Cattell. Scalable SQL and NoSQL Data Stores. *SIGMOD Rec.*, 39(4):12–27, May 2011.

[4] M. Chris and J. Zitting. *Tika in Action*. Manning Publications Co., 2011.

[5] W. S. Cleveland. *Visualizing data*. Hobart Press, 1993.

[6] J. Demšar, T. Curk, A. Erjavec, v. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14(1):2349–2353, Jan 2013.

[7] B. Fry. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. O'Reilly Media, Inc., 2007.

[8] G. Ghidini, S. Das, and V. Gupta. FuseViz: A Framework for Web-based Data Fusion and Visualization in Smart Environments. In *Proc. of IEEE MASS '12*, pages 468–472, Oct 2012.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov 2009.

[10] S. Jokić, S. Krco, J. Vuckovic, N. Gligoric, and D. Drajic. Evaluation of an XML database based Resource Directory performance. In *Proc. of TELFOR '11*, pages 542–545, Nov 2011.

[11] B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *Proc. of JCDL '09*, pages 313–322, 2009.

[12] B. Kules and B. Shneiderman. Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management*, 44(2):463–484, Mar 2008.

[13] K. K.-Y. Lee, W.-C. Tang, and K.-S. Choi. Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage. *Computer Methods and Programs in Biomedicine*, 110(1):99–109, Apr 2013.

[14] A. G. McQuillan. Honesty and foresight in computer visualizations. *Journal of forestry*, 96(6):15–16, Jun 1998.

[15] M. Paradies, S. Malaika, M. Nicola, and K. Xie. Comparing xml processing performance in middleware and database: A case study. In *Proc. of Middleware Conference Industrial Track '10*, pages 35–39, 2010.

[16] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava. Biketastic: Sensing and Mapping for Better Biking. In *Proc. of SIGCHI '10*, pages 1817–1820, 2010.

[17] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp. Fast and intuitive clustering of web documents. In *Proc. of KDD '97*, pages 287–290, 1997.

[18] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proc. of SIGIR '04*, pages 210–217, 2004.

Visualizing criminal networks reconstructed from mobile phone records

Emilio Ferrara
School of Informatics and Computing
Indiana University Bloomington, USA
ferrarae@indiana.edu

Pasquale De Meo
Department of Ancient and Modern Civilizations
University of Messina, Italy
pdemeo@unime.it

Salvatore Catanese, Giacomo Fiumara
Department of Mathematics and Computer Science
University of Messina, Italy
{scatanese,gfiumara}@unime.it

ABSTRACT

In the fight against the racketeering and terrorism, knowledge about the structure and the organization of criminal networks is of fundamental importance for both the investigations and the development of efficient strategies to prevent and restrain crimes. Intelligence agencies exploit information obtained from the analysis of large amounts of heterogeneous data deriving from various informative sources including the records of phone traffic, the social networks, surveillance data, interview data, experiential police data, and police intelligence files, to acquire knowledge about criminal networks and initiate accurate and destabilizing actions. In this context, visual representation techniques coordinate the exploration of the structure of the network together with the metrics of social network analysis. Nevertheless, the utility of visualization tools may become limited when the dimension and the complexity of the system under analysis grow beyond certain terms. In this paper we show how we employ some interactive visualization techniques to represent criminal and terrorist networks reconstructed from phone traffic data, namely foci, fisheye and geo-mapping network layouts. These methods allow the exploration of the network through animated transitions among visualization models and local enlargement techniques in order to improve the comprehension of interesting areas. By combining the features of the various visualization models it is possible to gain substantial enhancements with respect to classic visualization models, often unreadable in those cases of great complexity of the network.

Categories and Subject Descriptors

[**Information systems**]: World Wide Web—*Social networks*; [**Networks**]: Network types—*Social media networks*

Keywords

Mobile phone networks, criminal networks, visualization

1. INTRODUCTION

Criminal Network Analysis allows to identify structure and flow of information among the members of a criminal network and to acquire the knowledge necessary to plan proactive and reactive interventions. Among the more frequently used analytic techniques there is the mapping of interactions among the members of the organization and their activities by means of a graph [25]. A graph representation allows to overview the network structure, to identify the cliques, the groups, and the key players. The possibility of mapping the attributes of data and metrics of the network using visual properties of the nodes and edges makes this technique a powerful investigative tool. Often, however, visualization techniques become discouraging as a consequence of density and dimensions of the network. Some obstacles such as the overlap of nodes and the dense intersections of edges severely reduce the readability of the graph. In other words, there is a limit to the number of elements which can be distinctly viewed from the human eye. An influential theory about the improvement of the quality of network visualization has been suggested by Shneiderman in [31], where the so-called "Network Nirvana" is described. According to this theory, some demanding targets must be pursued: i) the visibility of each node; ii) the possibility of counting the degree of each node; iii) the possibility of following each edge from the source to the destination nodes and, iv) the possibility of identifying the clusters. Although it can be challenging, or even impossible, to satisfy all these conditions at the same time as the network grows in size and complexity, an effective network analysis strategy should try to optimize the visualization methods in order to incorporate these guidelines. In this work, we present three visualization techniques that yield better network representations that, in turn, allow for enhanced data interpretability; we discuss these layout techniques, namely fisheye, foci and network geo-mapping, specifically in the context of criminal network analysis, but we do not exclude a broader applicability to other domains of social network analysis (SNA).

1.1 Literature on criminal network analysis

In the last thirty years academic research related to the application of social network analysis to intelligence and study of criminal organizations has constantly grown. One of the most important studies is due to Malcolm Sparrow [33], related to the application of the techniques of analysis of networks, and their vulnerabilities, for intelligence scopes.

Sparrow defined four features peculiar of criminal networks (CNs), namely: i) limited dimension — CNs are often composed of at most few thousand nodes; ii) information incompleteness — criminal or terrorist networks are unavoidably incomplete due to fragmentary available information and erroneous information; iii) undefined borders — it is difficult to determine all the relations of a node; and, iv) dynamics — new connections imply a constant evolution of the structure of the network.

Thanks to Sparrow's work, other authors tried to study criminal networks using the tools of SNA. For example, Baker and Faulkner [3] studied illegal networks in the field of electric plants and Klerks [21] focused on criminal organizations in The Netherlands. In 2001, Silke [32] and Brennan et al. [8] acknowledged a slow growth in the fight against terrorism, and examined the state of the art in the field of criminal network analysis.

Arquilla and Ronfeldt [1] summarize prior research by introducing the concept of Netwar and its applicability to terrorism. They illustrate the difference between social networks and CNs, demonstrating the great utility of network models to understand the nature of criminal organizations.

All these early studies somehow neglected the importance of network visualization, stressing aspects related more to statistical network characterization, or interpretation of individuals' roles rooted in social theory. However, in 2006, a popular work by Valdis Krebs [22] applied graph analysis in conjunction with network visualization theory to analyze the Al Qaeda cell responsible of the 2001-09-11 terrorist attacks in the USA. This work represents a starting point of a series of academic papers in which social network analysis methods become applied to a real-world cases, differently from previous work where mostly toy models and fictitious networks were used. Krebs' paper is one of the more cited papers in the field of application of social network analysis to Criminal Networks and it inspired further research in network visualization for the design and development of better SNA tools applications to support intelligence agencies in the fight against terror, and law enforcement agencies in their quest fighting crime.

2. THE PROBLEM

In criminology and research on terrorism, SNA has been proved a powerful tool to learn the structure of a criminal organization. It allows analysts to understand the structural relevance of single actors and the relations among members, when regarded as individuals or members of (one or more) subgroup(s). SNA defines the key concepts to characterize network structure and roles, such as centrality [16], node and edge betweenness [16, 6, 14], and structural similarity [24]. The understanding of network structure derived from these concepts would not be possible otherwise [35]. The above-mentioned structural properties are heavily employed to visually represent social and criminal networks as a support decision-making processes.

SNA provides key techniques including the possibility to detect clusters, identify the most important actors and their roles and unveil interactions through various graphical representation methodologies [40]. Some of these methods are explicitly designed to identify groups within the network, while others have been developed to show social positions of group members. The most common graphical layouts have historically been the node-link and the matrix representations [17].

Visualization has become increasingly important to gain information about the structure and the dynamics of social networks: since the introduction of sociograms, it appeared clear that a deep understanding of a social network was not achievable only through some statistical network characterization [35].

For all these reasons, a number of different challenges in network visualization have been proposed [30]. The study of network visualization focuses on the solution of the problems related to clarity and scalability of the methods of automatic representation. The development of a visualization system exploits various technologies and faces some fundamental aspects such as: i) the choice of the layout; ii) the exploration dynamics; and, iii) the interactivity modes introduced to reduce the visual complexity.

Recent studies tried to improve the exploration of networks by adding views, user interface techniques and modes of interaction more advanced than the conventional node-link and force-directed [18] layouts. For example, in *SocialAction* [27] users are able to classify and filter the nodes of the network according to the values of their statistical properties. In *MatrixExplorer* [20] the node-link layout is integrated with the matrix layout. Nonetheless, these visualization systems have not been explicitly developed with the aim of the exhaustive comprehension of all properties of the network. Users need to synthesize the results coming from some views and assemble metrics with the overall structure of the network.

Therefore, we believe that an efficient method to enhance the comprehension and the study of social networks, and in particular of criminal networks, is to provide a more explicit and effective node-link layout algorithm. This way, important insights could be obtained from a unique layout rather than from the synthesis derived from some different layouts.

We recently presented a framework, called *LogAnalysis* [9, 15], that incorporates various features of social network analysis tools, but explicitly designed to handle criminal networks reconstructed from phone call interactions. This framework allows to visualize and analyze the phone traffic of a criminal network by integrating the node-link layout representation together with the navigation techniques of zooming and focusing and contextualizing. The reduction of the visual complexity is obtained by using hierarchical clustering algorithms. In this paper we discuss three new network layout methods that have been recently introduced in *LogAnalysis*, namely fisheye, foci and geo-mapping, and we explain how these methods help investigators and law enforcement agents in their quest to fight crime.

It's worth noting that various tools to support network analysis exist. However, only few of them have been developed specifically for criminal network investigations. We mention, among others, commercial tools like COPLINK [10, 37], An-

alyst's Notebook[1], Xanalysis Link Explorer[2] and Palantir Government[3]. Other prototypes described in academic papers include Sandbox [36] and POLESTAR [28]. Some of these tools show similar features to *LogAnalysis*, but, to the best of our knowledge, none of them yields the same effective and scalable network visualization with support to criminal networks reconstructed from phone call records.

2.1 Aspects of structural analysis

A central node of a criminal network may play a key role by acting as a leader, issuing orders, providing regulations or by effectively assuring the flow of information through the various components of the CN. The removal of these central nodes may efficiently fragment the organization and interrupt the prosecution of a criminal activity.

Apart from studying the roles of various members, investigative officers must pay particular attention to subgroups or gangs each of which may be in charge of specific tasks. Members of the organization must interact and cooperate in order to accomplish their illicit activities. Therefore, the detection of subgroups whose members are tightly interrelated may increase the comprehension of the organization of the CN. Moreover, groups may interact according to certain schemes. For example, the members of a clan could frequently interact with the members of another and seldom with the remaining members of the network. The detection of interaction models and the relations among the subgroups highlights information particularly useful about the overall structure of the network.

A significant aspect of the analysis of criminal networks is that it requires, differently from other networks, the ability of integrating information deriving from other sources in order to precisely understand its structure, operation and flow of information. A typical process employed by an investigator is to start from one, or a few, known entities; after analyzing the associations these entities have with others, if any interesting association emerges, one may follow such a lead and keep expanding the associations until any significant link is uncovered between seemingly unrelated entities.

Mobile phone networks and online platforms are constantly used to perform or coordinate criminal activities [38, 26]. Phone networks can be used to connect individuals involved in criminal activities in real time, often during real-world criminal events, from simple robberies to terror attacks. Online platforms, instead, can be exploited to carry out illicit activities such as frauds, identity thefts or to access classified information.

The analysis of a criminal network is thus aimed at uncover the structural schemes of the organization, its operations and, even more importantly, the flow of communications among its members. In modern investigative techniques the analysis of phone records represents a first approach that precedes a more refined scrutiny covering financial transactions and interpersonal relations. For these reasons a structured approach is needed.

[1] ibm.com/software/products/analysts-notebook/
[2] http://www.xanalys.com/products/link-explorer/
[3] http://www.palantir.com/solutions/

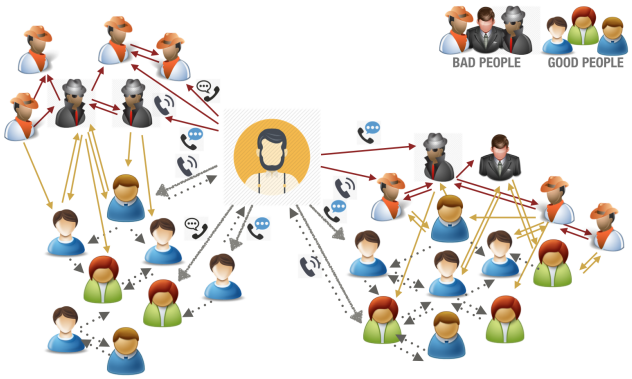


Figure 1: Phone calls network of a suspected. Investigators start from some known entities, analyze the associations they have with others and expanding the associations until some significant link is uncovered. Here are highlighted personal interactions (gray arrows), links between criminal and personal connections of the suspect (yellow) and connections between members of the organization (in red).

Figure 1 shows a stylized representation of a criminal network reconstructed from phone call records. We show the flow of phone communications of an individual subject of investigation, and we highlight various kind of phone interactions among individuals belonging to that person's social circles, and those belonging to the same criminal organization the individual is part of.

In the following we discuss three techniques that allow to efficiently and scalably inspect criminal networks reconstructed from phone interactions.

3. VISUALIZATION TECHNIQUES

Typical network visualization tools rely on the popular force-directed layout [18]. The force-directed model represents the structure of the graph on the same foot as a physical system, in which nodes are physical points subject to various forces; nodes' coordinates (and therefore the layout itself) derive from the search of an equilibrium configuration of the physical system modeled by the algorithm [7]. This particular layout arrangement has the advantage of grouping users in clusters which can be identified according to the heightened connectivity. The Barnes-Hut algorithm [4] associated to this layout simulates a repulsive N-body system in order to continuously update the position of the elements.

To optimize the visualization, it is possible to interactively modify the parameters relative to the tension of the springs (edges). Nodes with low degree are associated a small tension and the elements are located in peripheral positions with respect to high degree nodes. Other parameters can be tuned, such as spring tension, gravitational force and viscosity. Our goal, in the following, is to suggest two methods to improve force-directed based layouts. As we will show, these techniques are especially well suited for criminal network analysis; however, they could potentially be generalized for broader usage in other domains of network analysis — for example, for applications in social and political sciences.

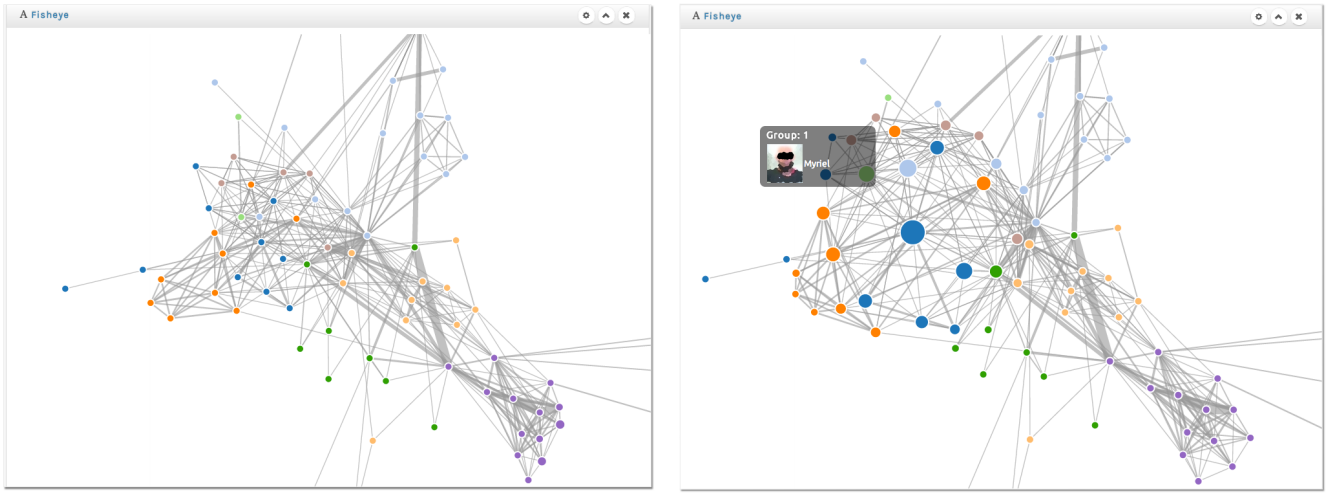


Figure 2: The left picture shows a force-directed layout of a criminal network. On the right we depict the fisheye view of the same graph using transformation with distortion.

3.1 Focus and context based visualization

The number of edges within a network usually grows faster than the number of nodes. As a consequence, the network layout would necessarily contain groups of nodes in which some local details would easily become unreadable because of density and overlap of the edges. As the size and complexity of the network grow, eventually nodes and edges become indistinguishable. This problem is known as visual overload [2]. A commonly used technique to work around visual overload consists of employing a zoom-in function able to enlarge the part of the graph of interest. The drawback of this operation is the detriment of the visualization of the global structure which, during the zooming, would not be displayed. However, such a compromise is reasonable in a number of situations including, in some cases, the domain of criminal network analysis.

During an investigation, it is crucial to narrow down the analysis to the relevant suspects, to efficiently employ human and computational resources. Police officers typically draw some hypotheses about an individual suspect of being part of a criminal organization, or of being involved (or about to) in some crime; they concentrate the initial investigation on this individual, and on that person's social circles, as a ground to build the social network object of analysis. The main role of visual analysis lies in allowing the detection of unknown relations, on the base of the available limited information. A typical procedure starts from known entities, to analyze the relations with other subjects and continue to expand the network inspecting first the edges appearing the most between individuals apparently unrelated. During this procedure, only some nodes are relevant and it is important to focus on them rather than on the network as a whole.

Nevertheless, a spring embedded layout (including force-directed ones) does not provide any support to this kind of focus and analysis. In these situations, *focus and context* visualization techniques are needed in order to help a user to explore a specific part of a complex network. To this purpose, we here introduce the fisheye and the foci layouts.

3.2 Fisheye layout

Focus and context is an interactive visualization technique [23]. It allows the user to focus on one or more areas of a social network, to dynamically tune the layout as a function of the focus, and to improve the visualization of the neighboring context. The *fisheye view* is a particular focus and context visualization technique which has been applied to visualize self-organizing maps in the Web surfing [39]. It was first proposed by Furnas [19] and successively enriched by Brown et al. [29]. It is known as a visualization technique that introduces distortion in the displayed information.

The fisheye layout is a local linear enlargement technique that, without modifying the size of the visualization canvas, allows to enhance the region surrounding the focus, while compressing the remote neighboring regions. The overall structure of the network is nevertheless maintained. An example of application of this technique is show in Figure 2. The picture shows a moderately small criminal network reconstructed from phone call interactions of about 75 individuals. The layout on the left panel is obtained by using a force-directed method implemented in our framework, *LogAnalysis*. The analyst can inspect the nodes of the network, which contains known criminals, suspects, and their social circles. When the focus is applied on a given node, the visualization transitions to the fisheye layout (see the right panel). A tool-tip with additional information about the node appears when the node is selected — it shows the phone number, personal details, address, photo, etc. The layout causes edges among remote nodes to experience stronger distortions than local nodes. The upside of the presented method is the possibility to achieve the three recommendations of Network Nirvana [30] when focusing on a given node: all the nodes' neighbors are clearly visible, the node degree is easily countable, and the edges incident on that node can be identified and followed.

Note that fisheye and force-directed layouts can be used in a complementary way. By combining the two methods, our framework efficiently yields focus and context views.

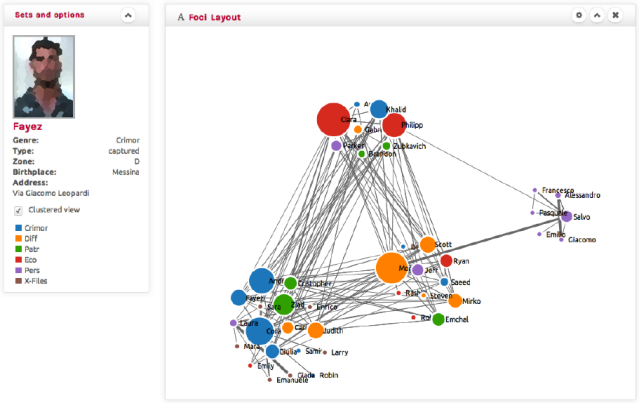


Figure 3: Foci layout.

3.3 Foci layout

The *foci layout* implements three network visualization models: force-directed, semantic and clustered layouts. The latter is based on the Louvain community detection algorithm [5, 11]. Future implementations will explore other methods [12, 13]. Our model supports multilayer analysis of the network through interactive transitions from the force-directed layout, with a single gravitational center, to the clustered one with more force centers placed in predetermined distinct areas. This layout allows to analyze the network on various layering levels depending on specified node attributes. Figure 3 shows the phone traffic network of some clans the previous criminal network, in which the color of the nodes denotes the type of crime committed by the members.

In this example, the clustering truthfully reflects the known territorial division among the groups belonging to the organization. In Figure 3 the focus is on a specific node. Using this layout it is possible to contextually analyze the community structure, the type of committed crime in respect to the members of the clan, and the direct relations of each single individual. This layout integrates also the forth Network Nirvana recommendation, namely the possibility to identify clusters and to highlight the community structure.

3.4 Network geo-mapping

It is possible to extend the phone traffic analysis to include the phone logs recorded by the BTS (Base Transceiver Station), in which the GPS coordinates of the cell are reported. All base stations are provided with directional antennas and each cell has two or more sectors. For each cell it is known the azimuth (direction) corresponding to the central axis of each sector, together with the width of the beam of each antenna, which determines the coverage angle of the sector. These data do not allow to localize the geo-referenced position of the phones involved in the events recorded in the logs. Nevertheless, it is possible, within a certain approximation, to localize the users falling within the coverage area.

Zang et al. [41] described a technique based on Bayesian interference to localize mobile phones using additional information, such as the round-trip-time of data transmission packets and the measure of SINR (Signal to Interference plus Noise Ratio). The parameters obtained experimentally have been compared with the records of phone calls and the corresponding GPS entries to ascertain their distribution. This localization technique produces satisfactory results with a reduction of the error amounting to a 20% with respect to the *blind approach*. Traag et al. in [34] used Bayesian interference to deduce, starting from phone traffic data, profiles about the places and the proximity of a given social event.

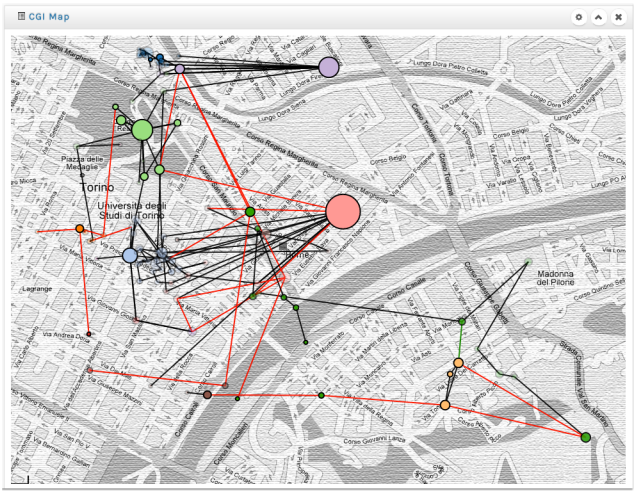


Figure 4: Geo-mapping layout.

Our framework provides network geo-mapping by using this type of techniques to infer the spatial origin of each call. We here describe the network geo-mapping visualization method adopted in *LogAnalysis*. This layout allows to simultaneously carry out spatial and temporal relational analysis of phone call logs. It places nodes of the network on a map, in correspondence of the coordinates of the cells linked during the events recorded in the logs. Nodes are connected by links related to displacements. Contacts falling within the sectors of a given zone are represented with nodes of the same color. Information about displacements, routines and areas of interest for the investigation are displayed. The adoption of network geo-mapping has proved extremely useful during real investigations. Figure 4 shows, as an example, a case study in which larger nodes identify zones in which, in the time period of the investigation, a high number of contacts has been recorded among some members of the CN. Unsurprisingly, the inspection by police officers of such high-profile locations provided crucial insights on the investigation. Unfolding the temporal evolution of the geo-mapped phone traffic network also allows to reproduce individuals' movements and communication dynamics during specific criminal events embedded in space and time, like robberies, assaults, or homicides.

4. CONCLUSIONS

Criminal network analysis benefits from visualization methods used to support the investigations, especially when dealing with networks reconstructed from heterogeneous data sources, characterized by increasing size and complexity. In this paper we integrated the spring embedded algorithm with the fisheye and foci layouts to allow interactive exploration of criminal networks through our network analysis framework. The combination of these techniques proved

helpful to support investigators in the extraction of useful information and critical insights, to identify key members in terrorist groups, and to discover specific paths of interaction among members of criminal organizations. Experimental results show that the combination of force-directed layouts, distortion techniques and multi-force systems yield better performance in terms of both efficiency and efficacy.

5. REFERENCES

[1] J. Arquilla and D. Ronfeldt. Networks and netwars: The future of terror, crime, and militancy. *Survival*, 44(2):175–176, 2001.

[2] J. Assa, D. Cohen-Or, and T. Milo. Displaying data in multidimensional relevance space with 2d visualization maps. In *Proc. Visualization '97*, pages 127–134, 1997.

[3] W. Baker and R. Faulkner. The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. *Am. Social. Rev.*, 58, 1993.

[4] J. Barnes and P. Hut. A hierarchical o(n log n) force calculation algorithm. *Nature*, 324:446–449, 1986.

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

[6] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[7] U. Brandes. Drawing on physical analogies. In *Drawing Graphs*, pages 71–86. Springer, 2001.

[8] D. W. Brannan, P. F. Esler, and N. T. Anders Strindberg. Talking to terrorists: Towards an independent analytical framework for the study of violent substate activism. *Studies in Conflict and Terrorism*, 24(1):3–24, 2001.

[9] S. Catanese, E. Ferrara, and G. Fiumara. Forensic analysis of phone call networks. *Social Network Analysis and Mining*, 3(1):15–33, 2013.

[10] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, and J. Schroeder. Coplink: managing law enforcement data and knowledge. *Comm. ACM*, 46(1):28–34, 2003.

[11] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Generalized Louvain method for community detection in large networks. In *Proc. 11th International Conference on Intelligent Systems Design and Applications*, pages 88–93. IEEE, 2011.

[12] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Enhancing community detection using a network weighting strategy. *Information Sciences*, 222:648–668, 2013.

[13] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Mixing local and global information for community detection in large networks. *Journal of Computer and System Sciences*, 80(1):72–87, 2014.

[14] P. De Meo, E. Ferrara, G. Fiumara, and A. Ricciardello. A novel measure of edge centrality in social networks. *Knowl-based Syst*, 30:136–150, 2012.

[15] E. Ferrara, P. De Meo, S. Catanese, and G. Fiumara. Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications*, 41(13):5733–5750, 2014.

[16] L. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[17] L. C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1, 2000.

[18] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[19] G. W. Furnas. Generalized fisheye views. *SIGCHI Bull.*, 17(4):16–23, Apr. 1986.

[20] N. Henry and J.-D. Fekete. Matrixexplorer: A dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):677–684, Sept. 2006.

[21] P. Klerks and E. Smeets. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? *Connections*, 24:53–65, 2001.

[22] V. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[23] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact.*, 1(2):126–160, June 1994.

[24] F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80, 1971.

[25] J. Mena. *Investigative Data Mining for Security and Criminal Detection*. Butterworth-Heinemann, 2003.

[26] C. Morselli. Assessing vulnerable and strategic positions in a criminal network. *Journal of Contemporary Criminal Justice*, 26(4):382–392, 2010.

[27] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Trans. Visual. and Computer Graphics*, pages 693–700, 2006.

[28] N. J. Pioch and J. O. Everett. Polestar: collaborative knowledge management and sensemaking tools for intelligence analysts. In *Proc. 15th ACM international conference on Information and knowledge management*, pages 513–521. ACM, 2006.

[29] M. Sarkar and M. H. Brown. Graphical fisheye views. *Comm. ACM*, 37(12):73–84, 1994.

[30] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *Proc. 9th SIGCOMM conference on Internet measurement conference*, pages 35–48. ACM, 2009.

[31] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Trans. Visual. and Computer Graphics*, 12(5):733–740, Sept 2006.

[32] A. Slike. The devil you know: Continuing problems with research on terrorism. *Terrorism and Political Violence*, 13:1–14, 2001.

[33] M. K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, 13(3):251–274, 1991.

[34] V. A. Traag, A. Browet, F. Calabrese, and F. Morlot. Social event detection in massive mobile phone data using probabilistic location inference. In *2011 IEEE 3rd international conference on social computing (socialcom)*, pages 625–628. IEEE, 2011.

[35] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge Univ. Pr., 1994.

[36] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: concepts and methods. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 801–810, 2006.

[37] J. Xu and H. Chen. Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Trans. on Information Systems*, 23(2):201–226, 2005.

[38] J. Xu and H. Chen. Criminal network analysis and visualization. *Comm. ACM*, 48(6):100–107, 2005.

[39] C. Yang, H. Chen, and K. Hong. Visualization of large category map for internet browsing. *Decis. Support Syst.*, 35(1):89–102, Apr. 2003.

[40] C. Yang, N. Liu, and M. Sageman. Analyzing the terrorist social networks with visualization tools. In *Intelligence & security informatics*. 2006.

[41] H. Zang, F. Baccelli, and J. Bolot. Bayesian inference for localization in cellular networks. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.

Towards Visual Overviews for Open Government Data

Alvaro Graves
Inria Chile
Av. Apoquindo 2827, piso 12
Santiago, Chile
alvaro.graves@inria.cl

Javier Bustos-Jiménez
NIC Chile Research Labs
Blanco Encalada 1975
Santiago, Chile
jbustos@niclabs.cl

ABSTRACT

The rise of Open Data initiatives has led to the publication of many datasets from different organizations and governments. These datasets cover a wide range of knowledge domains, from budget to education to health care. However, not all datasets have the quality, granularity or type of information that is relevant to each user. Moreover, in many cases the description or metadata does not specify clearly the content of a dataset, difficulting the exploration of datasets by stakeholders. In this paper we propose the use of dashboards and visualizations as a way to preview the content of datasets for easier exploration. The use of visualizations can provide a rapid way to select or discard datasets based on their content, reducing the potential datasets that a user may need to look in order to get what she needs.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques—*User interfaces*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Methodology*; I.7.4 [**Document and Text Processing**]: Electronic Publishing

General Terms

Documentation,Human Factors,Open Government Data

Keywords

Open Government Data, Open Data, Preview, Data Visualization

1. INTRODUCTION

Over one million datasets [16] are currently available in different portals across the globe. Although the data is publicly available, their organization and structure is not clear for all the stakeholders necessarily. For example, at the time of this writing the search for "child obesity" in *Data.gov* and *Data.gov.uk* (the two largest Open Government Data portals) gives different results, as can be see in Figure 1: In

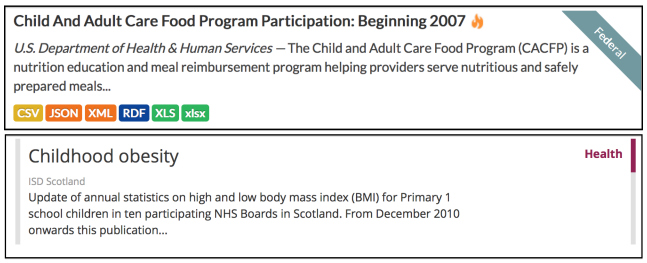


Figure 1: Restults provided by searching for `child obesity` in *Data.gov* (upper image) and *Data.gov.uk* (lower image)

Data.gov, only one dataset is available (in several formats). This dataset is described as "federal", however a closer look shows that the data is related to the state of New York only. In the case of *Data.gov.uk*, 16 results provide information related to child and obesity in PDF and Excel formats. Beyond these difference, it is not clear for a researcher or developer is these datasets are relevant to her needs; having a title and a description is useful, but does not clarify exactly what type of information, granularity and quality of the data is available.

For example, it is not clear what specific data is contained in a dataset, what structure is used or the scope of this dataset. As mentioned early,i n the case of the US dataset about child obesity, it is labeled as "federal", however the data describes only information about New York State; it is likely that other manually curated tags and descriptions may not be precise in terms of the content or scope of the datasets published. Thus, the question in this and many other cases is how can stakeholders know in advance what's in a dataset **before** downloading it? We propose the use of dashboards and visualizations to describe and preview the content of datasets; this visual representations will help stakeholder to decided whether a dataset is useful for them or not.

This paper is structured as follows: Section 2 describes related work found in the literature and state of the art technology. Section 3 discusses different pieces of information that can be used to create visual overviews from some of the more common file formats used to publish Open Government Data. In Section 4 we show a prototype developed as an example of what can be done to create visual overviews

of datasets using the information discussed previously. Section 5 presents the future challenges on our research and we discuss our conclusions in Section 6.

2. RELATED WORK

The problem of good data visualizations has been studied many years [24]. In terms of data exploration and visualization, Schneiderman [22] summarizes the Visual Information Seeking Mantra as *Overview first, zoom and filter, then details-on-demand*; humans need to get the "big picture" of a dataset first in order to decide where to explore next. Thus, a visual overview of a dataset can be be useful for researchers and journalists to know "what's in there" before taking further action.

One of the seminal works in dataset preview was made by Doan et al. in 1999. They studied the effects of visual previews of queries for NASA's EODIS datasets [19], concluding that the main advantages of these visual strategies were:

- "*eliminate zero-hit queries*,
- *reduces network activity and browsing effor by preventing the retrieval of undesired datasets*,
- *represents statistical information of database visually to aid comprehension and explorarion*,
- *support synamic queries, which aids users to discover dataset patterns and exceptions*, and
- (they are) *suitable to novice, intermittent, or expert users*".

A generalization of query previews is presented in the work of Tanin et al. [23], complementing the work of Doan et al. with barcharts in order to show data distribution.

In the beginning of this century, similar conclusions were reached by Green et al. in their study about how previews and overviews allow users to rapidly discriminate useful information from those not for interest [10], applying their findings in the interfaces provided by the Digital Library of Congress and concluding that "*previews should be available at a high level within a site so users get a taste of what is to come early in their visit*".

Nowadays, the principles behind above works seems to be suitable for open data publication, as it has been reported to be for web searching by the work of Dörk et al.[7], where they studied performance and benefits of a new approach called *visual exploration* for information seeking on the Web (Figure 2).

From the perspecting of the Open Government Data, visualizations are valuable and useful artifacts for users [12]; visualizations can provide feedback and help on the decision making process related to public policies. A survey [9] showed that many stakeholders found that users were interested in interacting with data via the use of visualizations. Hence, there is reasonable evidence to support our hypothesis that preview visualizations can be a useful tools for Open Government Data stakeholders.

Figure 2: Visual exploration interface proposed by Dörk et al.[7], which includes data collection choosers, visualization widgets, text query box and the current set of results.

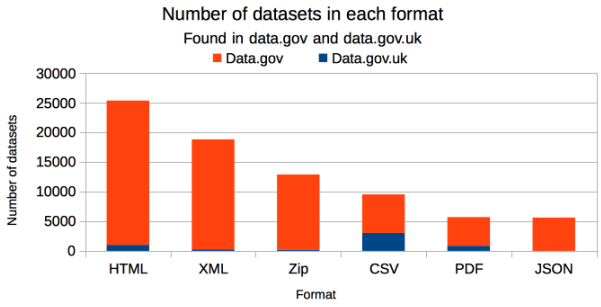


Figure 3: Number of datasets available in *data.gov* and *data.gov.uk* by format.

3. CONTENT FOR VISUAL OVERVIEWS

Different formats provide different support for data, metadata, annotations and other extra information that can be helpful for users to identify datasets that area valuable for them. In order to understand what format are more often used to publish Open Government Data, we looked at *Data.gov* and *Data.gov.uk*, two of the largest government data portals. We took the most popular formats reported by these portals and we found that most datasets are published in HTML, followed by XML, ZIP, CSV, PDF and JSON, as can be seen in Figure 3. It is important to note that in many cases a dataset is published in multiple formats, so these numbers are not related to the number of datasets available.

It is reasonable then to focus our efforts on the most common formats in order to cover an important number of datasets with our study. For this work, we do not considered ZIP files as part of the list of datasets to study, due to the fact that ZIP files are actually archives containing other files, such as CSV. Hence, for this study a ZIP file can be considered only as an "extra layer" of communication, and not a file format that we should study.

3.1 Data, metadata and annotations

We identify three different sources of information in a dataset that can be used to create visual overviews: data, metadata and annotations. We understand metadata different from annotations in that the former is aimed to provide machine-processable data about the dataset (e.g., creation date, author of the dataset), while the latter is more focused on explaining to a human reader certain aspects of the data (e.g., what does a field mean or information about how the data was collected). As mentioned before, different data formats provide different levels of support for data and metadata; thus, extracting data, metadata and annotations from different file formats present different challenges.

3.2 HTML

HTML is a markup language aimed to write "scientific documents, although its general design and adaptations over the years have enabled it to be used to describe a number of other types of documents" [25]. While not a data format *per se*, it has been widely used to publish data in a way that it is easy to consume by humans, via a web browser. There are multiple sources of data, metadata and annotations that we can use to represent visually.

- **Data:** Representing data in HTML can be done in multiple ways, from HTML tables to full web applications. In the most basic case, data can be presented as a list or a table, structured using the `<ul>`, `<ol>` or `<table>` elements. The process of extracting data from HTML documents is know as *Web Scraping* and there are many tools to do so [17][1]. This data can feed visual overviews to give insights about the actual content of the dataset.
- **Metadata:** HTML provides a mechanism to store metadata, by using the `<meta>` element. In the case of well-formed HTML tables, the header of these tables contain valuable metadata as well; the `<th>` element on a table will describe the name of each column, something that will tell a user if the dataset is useful for her purposes or not. These metadata elements can be extracted with web scraping techniques as well and used to give more insight about the structure of the data as well as more information about the provenance of it.
- **Annotations:** HTML supports comments in the code between <!– and –> strings sequences. These annotations can be used to extract information about the document and the data described in it as well. For example, it is possible to obtain the most relevant words in the comments and visualize them using a word cloud. It is important to note that in the case of HTML, many annotations might be related to the JavaScript code used in the document; a smart heuristic could discard potentially confusing annotations of this type.

3.3 XML

The *Extensible Markup Language* [4] is a language focused on structuring data for the Web, by providing a set of rules on how to encode such data. XML defines a tree-like structure where each node is a user-defined tag which may have content and attributes, as can be seen in Figure 4. There are several entities that can be extracted from a valid XML document to be used on a visual overview.

```
<persons>
 <person>
  <!-- this is a comment -->
  <name>John</name>
  <lastname>Doe</lastname>
  <language iso="EN">English</language>
 </person>
</persons>
```

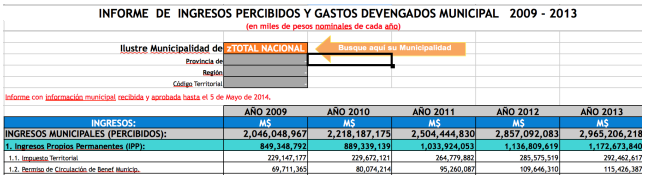
Figure 4: Example of a XML document.

- **Data:** It is possible to check for common words, numbers or phrases that occur in the content of XML tags. One way to do so is by using XPath [5], a query language aimed to extract data from XML documents. Similar to the case of HTML, the data can be used to inform the user about the actual content of the dataset.
- **Metadata:** There are at least two sources of information that can be used for a visual overview. First, the words used as tags and attributes are descriptive of the type of content that is about a dataset. For example in Figure 4, the words `person`, `name` and `lastname` give a good insight of what the data is about. Preprocessing the XML schema with Natural Language Processing techniques (e.g., Term frequency [20] or entity extraction [18]) can provide better insight on what type of information is contained in the dataset. Also, the structure how the data is organized is valuable in itself to understand the dataset; identifying the most common patterns in a XML structure and represent it visually, could give insight to users of what type of data is available, without the need to download the dataset.
- **Annotations:** XML allows comments in a similar way as in HTML (See Figure 4). XML Schema [14] also provides a series of non-mandatory mechanisms to annotate XML documents, by using the `xsd:annotation` tag. Applying NLP techniques as described above could help identify key entities related to this dataset (e.g., countries, contributors, organizations).

3.4 CSV

Comma-separated values is a loosely used term to define plain text files structured as tables, using separators (usually a comma, but semicolon and the *tab* character are not uncommon). CSV files are popular due to its simplicity in terms the readability and processing of the data, done both by humans and computers. In many cases, CSV are the result of exporting Spreadsheet files (such as Microsoft Excel) into text. In many cases it is possible to observe headers that defined the columns of a CSV file.

- **Data:** Since a CSV file is basically a table, it is possible to extract the most common terms found in the cells and display them as a bar chart or a word cloud



INFORME DE INGRESOS PERCIBIDOS Y GASTOS DEVENGADOS MUNICIPAL 2009 - 2013

(en miles de pesos nominales de cada año)

Ilustre Municipalidad de ¿TOTAL NACIONAL | Busque aquí su Municipalidad

Provincia de

Región

Código Territorial

Informe con información municipal recibida y aprobada hasta el 5 de Mayo de 2014.

INGRESOS:	AÑO 2009 M$	AÑO 2010 M$	AÑO 2011 M$	AÑO 2012 M$	AÑO 2013 M$
INGRESOS MUNICIPALES (PERCIBIDOS):	2,046,048,967	2,218,187,175	2,504,444,830	2,857,092,083	2,965,206,218
1. Ingresos Propios Permanentes (IPP):	849,348,792	889,339,139	1,033,924,053	1,136,809,619	1,172,673,840
1.1. Impuesto Territorial	229,147,177	229,672,121	264,779,882	285,575,519	292,462,617
1.2. Permiso de Circulación de Benef Municip.	69,711,365	80,074,214	95,260,087	109,646,310	115,426,387

Figure 5: Example of a spreadsheet version of a dataset. The CSV version does not respect the table structure, due to the titles and headers that are exported along with the rest of the data.

or other way to present it as a visual overview of the dataset. There are tools and libraries for virtually any programming language to read and extract data from CSV files.

- **Metadata:** Due to its simplicity, little metadata can be found in a CSV file. However, as mentioned before, in many cases CSV files contain headers that can be used to identify the topics described in the dataset.
- **Annotations:** CSV does not support annotations, however in many cases, the direct translation from a spreadsheet, such as Microsoft Excel, carries the title and other comments available on it (see Figure 5 as an example). These annotations break the table structure of the CSV file and makes it difficult to read it by programs. Still, these annotations can provide useful information about the content of the file. An heuristic to obtain such annotations could be the following: Read each line of a CSV file and consider it as an annotation, until the header is found.

3.5 PDF

The use of PDF files to publish data is a common practice among practically all governments and organizations, although it is widely discouraged and criticized [15][8]. One of the main reasons is that PDF is a *document* format, not a data format. In this sense, PDF does not comply with the Open Government Data principle [11] that states that data should be in a machine-processable format. Still, many efforts like Tabula [2] have been developed to extract data from PDF files.

- **Data:** As mentioned before, in the best of cases PDF files contain data tables that can be extracted semi-automatically to generate visualizations, similar to the case of CSV files.
- **Metadata:** Although PDF supports metadata and embeddable raw data [13], common tools for creating PDFs do no include metadata but some basic authorship information. It is not clear what type of metadata may be available in the general case to use for an visual overview.
- **Annotations:** Similar to the case of XML and HTML documents, annotations in PDF can be used to identify relevant terms that can be later used create a visual overview.

```
{
  persons: [
       {
         name: "John",
         lastname: "Doe",
         language: {
                     value: "English",
                     iso: "EN"
                   }
       }
  ]
}
```

Figure 6: A possible JSON representation of the data shown in Figure 4 as XML.

3.6 JSON

The JavaScript Object Notation JSON, is an open standard format that has gained popularity, especially in the Web development community, due to the simplicity for consumption by humans and machines alike. JSON provides a mechanism to transmit objects that can be use to communicate different types of variables. Many see JSON as a simpler, easier-to-use alternative to XML [6]. An example of a JSON document can be seen in Figure 6.

Similar to XML, JSON provides a tree-like structure, but supports different data types, arrays and other objects as well. Thus, it is possible to extract similar information as in the case of XML to later be visualized.

- **Data:** The values in a JSON document can be used to obtain the most significant words or phrases that can be used later to create a visualization.
- **Metadata:** Collecting the words used as keys can give insights on what type of data is presented in the document. Also, the tree structure could be used to identify how the data is modeled.
- **Annotations:** JSON does not provide a way to annotate or comment documents.

4. PROTOTYPE

As a way to test our ideas, we developed a demo tool that creates a visual overview of a dataset. This visual overview consist on a sample of the data and a dashboard based on the information extracted from a dataset. Due to simplicity, our prototype only works with CSV files, but the principles shown are the same for the other file formats described in Section 3. We implemented this demo using JavaScript and the D3.js library [3]. The prototype is available at `https://github.com/niclabs/visual-overview` as open source software.

4.1 Rationale

The prototype presents three different levels of detail of the data contained in a dataset. First, we considered useful to give the user a sample of the data, so she can get an idea of what it looks like as a table. To do so, we included the first three rows of the dataset.

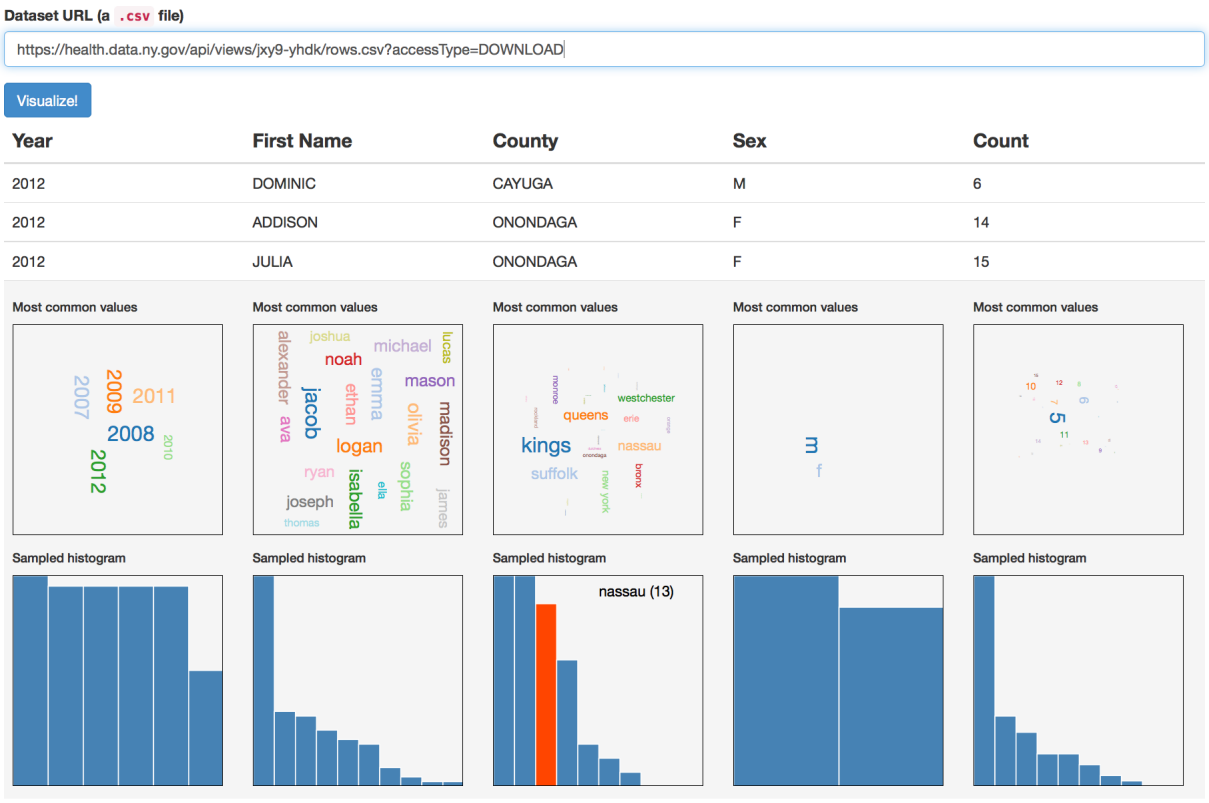


Figure 7: Screenshot of our prototype. A user can indicate a CSV file available and the system will render several statistics related to the values present in the data, as well as the headers available.

Second, in our experience most CSV files describe data properties in terms of columns (in contrast to rows); a CSV column usually contains values related to a specific dimension (e.g., age, latitude, name). Thus, one reasonable approach is to create visualizations for each column. As a way to provide a visual representation of the values on each column, we used *word clouds* [21]; in this way, we present the most common values in each column to the user in a way that is easy to consume without any technical background.

Finally, in many cases it is important to provide more information about the distribution of values to answer questions, such as *Is the data normally distributed? Does it follow a long tail? Are all the values equally likely?*. Although the word cloud provides some insights on this respect, we think a clearer representation was needed. Thus, a histogram of the values in each column is provided. This histogram facilitates the understanding of how the data is distributed and what are the most/least common values.

It is important to note that as a prototype, there are many issues with this software. For example, a more sophisticated approach would consider the type of data (i.e., generic numbers, strings, geographical coordinates, time and dates) and use different visual strategies that are more suitable for each case. The variety of the available values may also affect what visual strategy could be used; for example, for the columns `sex` and *count* the use of word clouds is not necessarily the best strategy.

4.2 Use of the prototype

After a user has entered the URL of a dataset, the prototype will analyze the data in order to extract the more common terms. Although our prototype processes the data live, it is possible to imagine more sophisticated mechanisms that deal with larger datasets, such as offline or batch processing. As mentioned early, our prototype provides several visualizations for each column of the CSV file, including data sample, a wordcloud and a histogram for each column. A screenshot of our prototype can be seen in Figure 7.

5. FUTURE WORK

Our hypothesis is that these visualizations can facilitate the process of deciding if a dataset is useful for a person or not. Thus, we propose to perform a user study to evaluate how easy or hard is for a user to find valuable information in the presence/absence of visual overviews. Also, the effectiveness of visual overviews may also depend on the type of visualizations that are displayed in different scenarios. Further research is necessary in this regard.

From this prototype, we can also take several paths. We plan to include support for other data formats, as described in Section 3. Having a web-based service available to preview and give insights about a dataset can be a valuable tool for journalists, activists and Open Government Data

researchers in general. Another option is to promote the use of tools similar to our prototype to be part of government data portals by default. Most of government organizations already provide a series of tags to help people identify and understand what each dataset is about. Adding an visual overview will help them on that effort. Finally, a smarter set of heuristics could be included in our prototype to provide more suitable visual representations, based on the type of data available in each dataset. Also, the use of annotations in datasets could be used to highlight certain visualizations over others.

6. CONCLUSIONS

In this paper we have proposed the use of visualizations to preview and give insights about datasets that can be useful and valuable to many stakeholders. We showed that for most of the more common file formats used to publish Open Government Data, it is possible to extract valuable information that can be later used to create visual overviews. We also showed how these visual overviews can be created using a prototype developed by the authors that present a dashboard of visualizations based on the information obtained from a dataset. Finally, we discussed the different paths this work can take in the future.

7. REFERENCES

[1] (2011) ScraperWiki. ScraperWiki.2011.http://scraperwiki.com/.

[2] "Tabula," http://tabula.nerdpower.org/, 2013.

[3] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[4] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (xml)," *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, 1998.

[5] J. Clark, S. DeRose *et al.*, "Xml path language (xpath) version 1.0," 1999.

[6] D. Crockford, "Json: The fat-free alternative to xml," in *Proc. of XML*, vol. 2006, 2006.

[7] M. Dörk, C. Williamson, and S. Carpendale, "Navigating tomorrow's web: From searching and browsing to visual exploration," *ACM Transactions on the Web (TWEB)*, vol. 6, no. 3, p. 13, 2012.

[8] M. Fioretti, "Open data: Emerging trends, issues and best practices," *Laboratory of*, 2011.

[9] A. Graves and J. Hendler, "Visualization tools for open government data," in *Proceedings of the 14th Annual International Conference on Digital Government Research*. ACM, 2013, pp. 136–145.

[10] S. Greene, G. Marchionini, C. Plaisant, and B. Shneiderman, "Previews and overviews in digital libraries: Designing surrogates to support visual information seeking," *Journal of the American Society for Information Science*, vol. 51, no. 4, pp. 380–393, 2000.

[11] O. G. W. Group *et al.*, "Principles of open government data," in *Workshop held in Sebastopol, CA, USA. http://www. opengovdata. org/home/8principles*, 8.

[12] J. Hoxha and A. Brahaj, "Open government data on the web: A semantic approach," in *Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on*. IEEE, 2011, pp. 107–113.

[13] King, James C., "Role of PDF and Open Data," in *Open Data on the Web, Campus London, Shoreditch, 2013*, 2013.

[14] A. Malhotra and P. Biron, "XML schema part 2: Datatypes," *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*, 2004.

[15] Manning, Nathaniel. (2013) Bad metrics and PDF graveyards: why development needs open data. http://www.theguardian.com/global-development-professionals-network/2013/oct/21/development-open-data-action.

[16] C. Peng. (2012, Aug.). int. open government data search data analytics. *linking open government data*. [Online]. Available: http://logd.tw.rpi.edu/iogds_data_analytics[RetrievedNov.24,2013]

[17] R. B. Penman, T. Baldwin, and D. Martinez, "Web scraping made simple with sitescraper," 2009.

[18] M. Pennacchiotti and P. Pantel, "Entity extraction via ensemble semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 238–247.

[19] C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns, "Interface and data architecture for query preview in networked information systems," *ACM Transactions on Information Systems (TOIS)*, vol. 17, no. 3, pp. 320–341, 1999.

[20] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1983.

[21] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer, "On the beauty and usability of tag clouds," in *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE, 2008, pp. 17–25.

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

[23] E. Tanin, C. Plaisant, and B. Shneiderman, "Broadening access to large online databases by generalizing query previews," *The craft of information visualization: readings and reflections*, p. 31, 2003.

[24] E. R. Tufte and P. Graves-Morris, *The visual display of quantitative information*. Graphics press Cheshire, CT, 1983, vol. 2.

[25] W3C, "HTML5, A vocabulary and associated APIs for HTML and XHTML," http://www.w3.org/TR/html5/introduction.html, 2014.

Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories*

James Honaker
Institute for Quantitative Social Science
Harvard University
1737 Cambridge St
Cambridge, MA 02138
jhonaker@iq.harvard.edu

Vito D'Orazio
Institute for Quantitative Social Science
Harvard University
1737 Cambridge St
Cambridge, MA 02138
dorazio@iq.harvard.edu

ABSTRACT

We detail our construction of *TwoRavens*, a graphical user interface for quantitative analysis that allows users at all levels of statistical expertise to explore their data, describe their substantive understanding of the data, and appropriately construct and interpret statistical models. The interface is a browser-based, thin client, with the data remaining in an online repository, and the statistical modeling occurring on a remote server. In our implementation, we integrate with tens of thousands of datasets from the *Dataverse* repository, and the large library of statistical models available in the *Zelig* package for the *R* statistical language. Our interface is entirely gesture-driven, and so easily used on tablets and phones. This, in combination with being browser-based, makes data exploration and quantitative reasoning easily portable to the classroom with minimal infrastructure or technology overhead.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical Software; H.4 [**Information Interfaces and Presentation**]: User Interfaces—*graphical user interfaces*; G.4 [**Mathematical Software**]: User Interfaces

Keywords

statistical user interfaces, open data, directed graphs

1. INTRODUCTION

With the proliferation of open data sources and replicable data repositories comes the promise of increased access to scientific information and the democratization of quantitative knowledge. However, meaningful analysis of this multitude of data remains outside the grasp of analysts who lack training in statistical modeling or knowledge of and access to statistical software platforms. For those with the necessary expertise, access is still limited by data transfer bottlenecks and installation and hardware overhead. To reduce barriers to statistical and quantitative reasoning and to promote the proliferation of empirical knowledge, we introduce the TwoRavens interface.

TwoRavens is a gesture-driven, Web-based device for the analysis of statistical models and the exploration of data. It is open-source and integrates with Dataverse repositories for open archiving of data[12, 6, 7], providing users access to the more than 50,000 data files currently housed in Dataverse repositories, as well as new data that users may upload free of charge. The statistical analysis component is powered by Zelig, a library for statistical inference in *R* that provides access to dozens of statistical models through a common call structure [14, 11]. TwoRavens links the power of Zelig to one of the largest database collections anywhere, requires no installation, and is accessible to devices ranging from mobile phones to tablets to smartboards.

For ease of use, the UI is designed to mirror the common quantitative workflow. Moving left to right on the interface, users make intuitive column and row selections or subsets of the data.[1] Users may perform transformations on these selections using the input box in the top right. In the center, users specify the intended statistical model in the framework of a directed graph, and have the option to perform transformations on selected columns or tag selections with specific properties, such as "nominal." Finally, on the right, users select their statistical model and have the option to set covariate values on right-hand side variables to obtain additional bootstrapped simulations of quantities of interest. Upon estimation, statistical results are displayed using a combination of graphs and tables.

In this article we describe the UI and its integration with Dataverse and Zelig. In future research we intend to build upon this foundation in two ways. First, as users supply information about hypothesized relationships in the data, TwoRavens will synthesize their input and automatically suggest appropriate modeling decisions, such as functional form and choice of model. Second, as models are estimated, the results will be cumulatively recorded across users to form a coherent map and meta analysis of the statistical results that exist in Dataverse.

2. DESIGN GOALS AND PRINCIPLES

In designing this software, some key goals were set that we believe the next generation of statistical tools will need to achieve. The challenges and solutions to reaching these goals informed some general design principles, which we feel are important to discuss and convey for future researchers. A paramount goal was to keep the interface as a **thin client**, keeping both the data itself, and the work of statistical

*All code, documentation, and numerous examples for our open source *TwoRavens* software project are available at: `http://datascience.iq.harvard.edu/tworavens`

[1]Depending on the discipline, columns are known as variables, fields, features, etc., and rows are known as observations, cases, units of analysis, etc.

processing, remote from the interface. We also required a **broadly accessible** platform that could be productively used by individuals at all levels of statistical expertise, from novice users with no statistical training, to experienced quantitative researchers. Furthermore, to make data accessible to users with different levels of available computational infrastructure, we worked to make this software **device independent** and feasible not only on traditional computers, which have previously been the sole domain of statistical software, but tablet, mobile, and other smart devices.

2.1 Goal: Thin Client

Although we have written an interface for statistical data analysis, neither the data itself, nor any statistical analysis, occurs or resides at the client side.[2] It is a challenge to explore data, and to implement statistical models, without the the immediacy of local interaction. However, having a thin client with remote data and remote statistical processing allows some enormous, novel advantages that are increasingly useful in modern data-intensive science: 1) When datasets are large, transferring the data to the user for exploration can be the primary bottleneck for speed, which this overcomes. 2) Similarly, when datasets are large, the final analysis might require distributed processing, which again requires another transfer of the data. In our architecture, again, the code for statistical processing is instead placed close to the original data. 3) When the data has privacy implications, there are settings where it is allowable to grant access to summaries and statistical representations of the data, (or privacy preserving versions of these statistics, such as those that are provably differentially private [9, 19, 8]), but not grant access to individual observations of the data itself. This architecture, where both the data and the analysis on the data remains remote, can be an enforceable way to grant access to private data in these settings.

2.2 Goal: Broadly Accessible

Statistical consulting is increasingly a necessary service provided by research universities to aid their research faculty. Most substantive researchers are heavily invested in the collection of their data, understand their variables well, and possess expert substantive knowledge about the plausibility of various relationships. What they may not know is the set of plausible statistical models appropriate to their objectives, the statistical language to describe their goals and aims, or the software knowledge to implement these choices. Consultants often provide that link, allowing experts without statistical training to translate their knowledge and goals into appropriate statistical software routines, and interpreting the results back in a substantively meaningful fashion. Obviously, however, not every user of data has access to a trained statistician. One overarching goal of the TwoRavens project is to provide as much of this service as can be automated. At this stage of the interface, this means providing a way for users to intuitively communicate all their knowledge about the data, so an appropriate model can be constructed for the quantitative task at hand.

2.3 Goal: Device Independent

Part of expanding the set of users who have access to quantitative reasoning, means lowering the infrastructure requirements to statistical analysis. While professional researchers may have extensive computational resources, many other settings, especially community colleges and high schools, have limited resources for instructors and none for students. Even free and open source statistical packages like *R* generally require the expenses of some level of IT support to install in a networked lab. Moreover, even in well-funded universities, substantive classes that use quantitative data may have no classroom lab access. Our goal is to enable a computationally capable access to statistical exploration with very minimal infrastructure.

2.4 Resulting Design Principles

In meeting these goals, we developed the following design principles for this interface: 1) **Browser Based:** The ability to run the software entirely through a browser, with no additional installed software, is possible because of the thin nature of the interface, and helps facilitate device independence as it can run on any device that includes a browser 2) **Gesture Driven:** Common statistical packages are normally command line, script, or menu driven, but to facilitate ease of use across devices and without user expertise, we have tailored a process that allows full exploration and setup of statistical models by interacting directly with the data though gesture 3) **Graphical Representation:** Directed graphs, also called probabilistic networks, are increasing used by statisticians as a representation of complex data generation processes [10], particularly in structural equation modeling as well as casual inference [16]. Using directed graphs to convey possible relationships between variables allows novices to convey all of their substantive understanding of a dataset, substantive experts and teachers to communicate in a manner conducive to qualitative discussion, and still allow statisticians to explicitly define their hypotheses of the data generating process. 4) **Maximal computational leverage:** Most statistical packages are interactive in a command driven relationship, waiting for instructions before creating any analysis product. To facilitate the architecture of the interface, we instead have an impatient design that preprocesses any graphs and summary statistics that we can envision might be needed, so they are already loaded on the thin client, without needing data interactions or queries. Much of this preprocessing can be done when the data is ingested to the repository, before the user has even discovered the data.

3. TWORAVENS SOFTWARE

The interface is written in Javascript and incorporates aspects of interactive graphics, Web-based statistics, and customized R applications. The interactive visualizations use Data-Driven Documents (D3), which has been influenced by tools such as Protovis and Processing [4, 3, 17]. Although components of TwoRavens appear as a Web interface for statistical software (comparable to R-fiddle or SAS OnDemand), its interaction with R is closer to that of applications created using tools such as RStudio's Shiny [18]. In its entirety, TwoRavens is comparable to GleamViz, which contains an interactive statistical modeling tool and whose remote servers handle the necessary processing [5]. Where TwoRavens distinguishes itself from GleamViz is that it does not require a desktop client and it is a general-purpose statis-

[2]Or slightly rephrased, we have built a package for statistical analysis of data, that can not handle data, nor analyze any statistics.

tical modeling tool, whereas GleamViz is tailored for modeling infectious diseases. Thus, as a Web-based, gesture-driven tool for statistical modeling, TwoRavens is unique.

Furthering its distinction from existing statistical software, it integrates with Dataverse, providing instant access to tens of thousands of datasets by simply launching TwoRavens from any Dataverse repository page, and with Zelig, delivering analysts meaningful yet easily interpretable statistical estimates and graphics. Given the availability of open access data in repositories such as Dataverse, the open source power of R, and the trend towards Web-based, interactive visualizations, such a device ties together many threads of modern quantitative computing.

3.1 Dataverse and Zelig Integration

In the last decade, the Data Science team at Harvard's Institute for Quantitative Social Science (IQSS) has developed software infrastructure and tools to facilitate and enhance data sharing, preservation, citation, reusability and analysis [13]. Over that time, the team has continuously developed two software products now widely used by the research community: *Dataverse*, a repository infrastructure for sharing research data, and *Zelig*, a statistical package for *R*.

Dataverse is "an open source data repository, which allows one to publish, share, reference, extract, and analyze research data" [7]. The Harvard Dataverse Network contains more than 52,000 studies with more than 700,000 files, and is the world's largest collection of social science data sets (`http://thedata.org`). The primary connection between TwoRavens and Dataverse is an API accessing metadata that complies with formatting standards set by the Data Document Initiative (DDI), an organization that promotes diligent data management [1]. In addition to keeping it thin, this facilitates the deployment of TwoRavens to any data repositories that comply with DDI standards.

Zelig is a wrapper and interface that allows a large body of different statistical models in the R statistical language to be used from a unified call structure [11, 14]. It is also a modeling architecture that interprets these statistical models in a substantively meaningful fashion [15]. By integrating with Zelig, TwoRavens has minimal backend manipulations, other than to map the information that has been entered by the user into an appropriate Zelig call. Thus, new functionality in Zelig may be translated to new functionality in TwoRavens rather seamlessly. Both R and Zelig are open source and freely available.

4. THE USER INTERFACE

Javascript is lightweight, and allows us to run the interface entirely through an internet browser. For ease of use and cross-platform portability, all functionality has gesture-driven capability, so statistical models on datasets stored in online repositories could be run from a tablet or mobile device without a keyboard. This ability expands the set of ways individuals use archived data and quantitative analysis, including bringing real-time data analysis into the classroom without using a computer laboratory setting.

A screenshot of TwoRavens is shown in figure 2. The workflow of developing a statistical analysis moves from left to right, first examining and selecting variables in the dataset, then constructing a framework of possible relationships, and then choosing and interpreting an appropriate statistical model for that framework. In figure 2, the user has selected a model, tagged `ti_cpi` as the dependent variable, and graphed an appropriate statistical model.

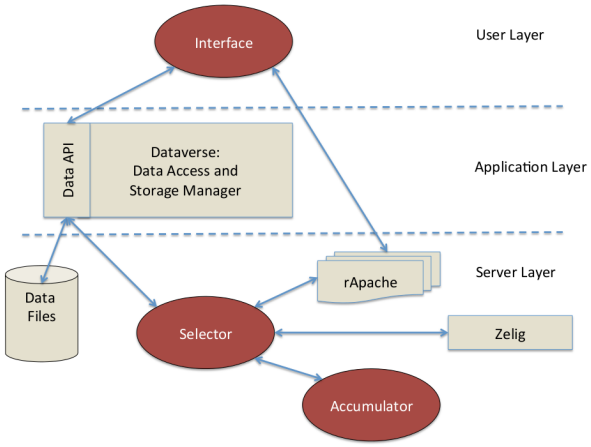


Figure 1: Architecture, including integration with *Dataverse* archival data repository, and *Zelig* library of statistical models for *R*.

4.1 Left Panel - Data Selection

Each variable in the dataset can be introduced as a node in the center space by clicking on the variable name in the left panel. Users may not be familiar with each variable in the dataset, so on mouseover information pertaining to that variable appears, such as primary summary statistics, graphs, metadata and short descriptions of the variable.[3] This information is precalculated by Dataverse when a new dataset is uploaded, and stored in a json file that is compatible with the DDI schema [1].

Although not explicitly in the left panel, TwoRavens includes an option to perform transformations on variables, such as taking the log of a variable or multiplying it by some factor, using the input box in the top right. Users may transform variables in two ways: (1) by manually entering text into the transformation input box; or (2) by clicking on the input box, selecting a variable from the drop down list, and then selecting a transformation from the function list. In this way, the gesture-driven functionality is preserved, while allowing additional flexibility for users familiar with *R* functions.

Users may be interested in subsetting the dataset to examine only observations that have specific values (for example, only European countries, or only respondents over the age of 60). For such cases, we include a *Subset* tab that shows the distribution of each variable. These distributions are either a density plot or a bar plot, depending on the variable's level of measurement and its number of unique values. By brushing the plots with the pointer, users select ranges of that variable upon which the data is to be subsetted. The numbers associated with the range are shown so that users may be precise in the ranges they specify. All metadata is remotely recalculated for the new subset using the same Dataverse ingest routines, and a new space that represents the subsetted data is added to the carousel in the center space of the interface.

[3]For devices not compatible with mouseover, we enable this feature via click and hold.

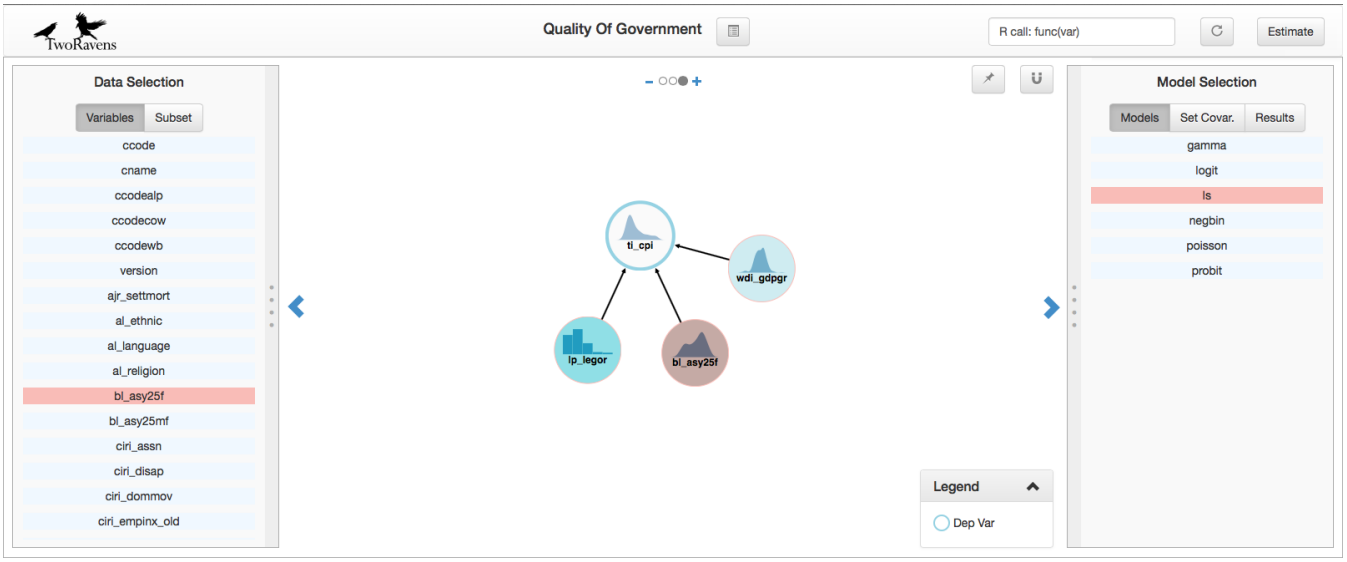


Figure 2: Basic UI with Dependent Variable and Model Selected

4.2 Center Space - Relationship Mapping

In the focal, central space, the substantive knowledge of the researcher is easily communicated by drawing a directed graph representation. With a two-finger click, arrows are drawn connecting nodes, depicting the possible relationships between variables. The nodes and arrows are an application of D3's force layout. In the simulated forces that propel the visualization, each node has a gravity, which draws all the nodes together, as well as a charge that repels them from getting too close. The arrows act as simple springs, and are removed when clicked. As the graph between the variables is built up by the researcher, it dynamically rearranges itself by these simulated forces. The researcher can also physically drag the pieces around, which acts as an additional force in the dynamic visualization, or completely turn off the simulated forces, and place every node manually for complete control of the representation (using the *force* toggle icon, represented by a pin).

Dependent variables, time, cross-sectional, and nominal identifiers are properties that may be tagged, by the user, to a node. Each property is denoted with their own colored halo. To tag a property, a user mouses over a node in the directed graph, at which point buttons appears as arcs around the perimeter of the node. Each arc is colored and labeled, and clicking on the arc associates that property to that node. At this point, the halo is colored appropriately and a legend appears, reminding users what the color of the halo represents. Additionally, the background color of the tagged variable in the left panel is changed to reflect the color of the tagged property.

The portion of the center space that is visible is actually just one element of a carousel, and users have the option to add and remove new workspaces (hereafter, elements). For example, if a user subsets the data, then an additional element is added to the carousel that corresponds to the subsetted data. Users toggle between elements by clicking a chevron or by clicking and swiping in the direction they want the carousel to move. Their current element is shown by highlighting its corresponding dot in the top-center of the center space. By clicking on the plus sign to the right of the dots, users are duplicating the current carousel, and placing its representation at the right of the element array. By clicking the minus sign, users are dropping this element from the carousel. To clear a modeling space but not drop it from the carousel, users may select the Erase icon, represented with a magnet.

4.3 Right Panel - Model Implementation

After examining the data and constructing a diagram of relationships, in the right panel users begin to investigate statistical models. The *Models* tab provides a list of the available statistical models that can be employed by Zelig. On mouseover, users see a brief description of the model so that they have some guidance as to which to select.

The next tab, labeled *Set Covar.*, provides the ability to interpret any estimated model by means of predicted and expected values at chosen values of the covariates, as well as first differences created by the changes in the predictions across changes in the covariates [15]. As shown in figure 3, users may choose values of the covariates at which to interpret the model by means of sliders superimposed on the densities of the variables. The slider positions initially default to the mean of each variable, while the scale of the slider marks each standard deviation away from the mean within the range of the variable. For bar plots, the value of the bar is also placed on the slider's scale.

When this information is complete, the researcher can estimate the model by clicking the *Estimate* button. At this time, the information extracted from the user is passed to an instance of an *R* application hosted on a remote server.[4] This remote application first builds a formula representation of the model from the graph connections the researcher has constructed. In the present version, this is all variables that have a path to the dependent variable, but more complex graphs can include intermediate or post-treatment variables as well as consequences of the dependent variable, which are

[4]We developed our *R* application in Rook and host it in rApache.

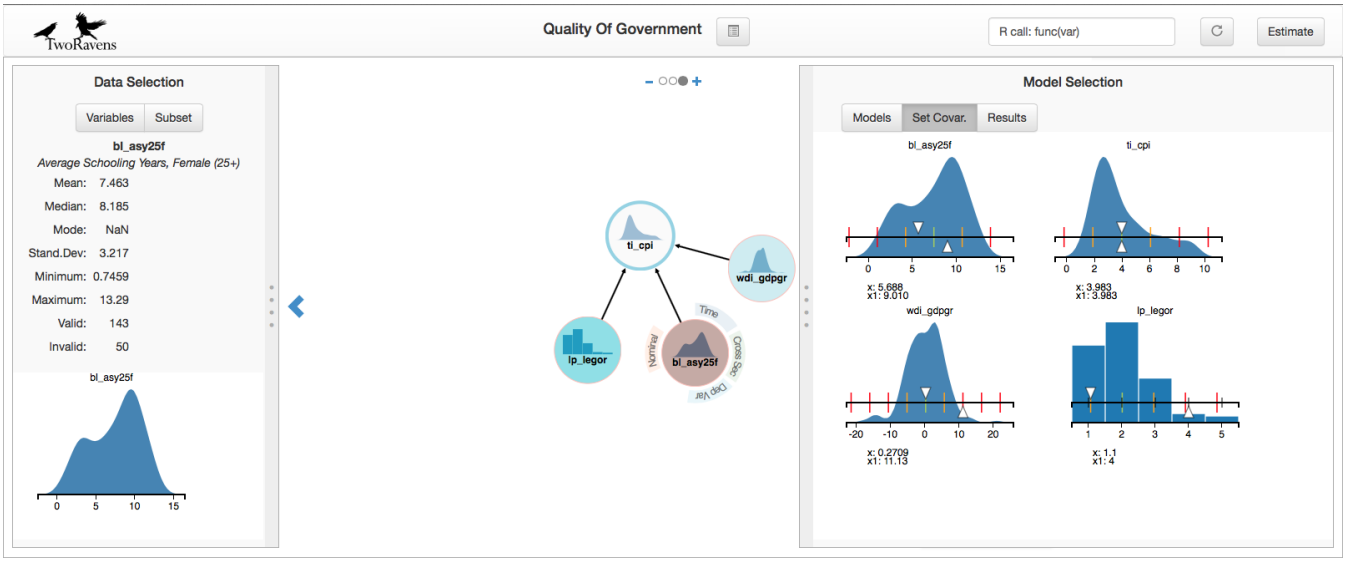


Figure 3: Node Description (left panel) and Set Covariate Values (right panel)

useful for forecasting and imputation, but omitted from the formula for a causally oriented analysis. Using this formula, the *R* application calls the applicable statistical model from the Zelig library. The results from Zelig are asynchronously returned to the browser interface. As can be seen in figure 4, the plots that Zelig produces, as well as a table of estimates, are available for viewing inside the *Results* tab.

5. EXAMPLE APPLICATION

The Quality of Government (QoG) represents one database for which the primary benefits of our tool might be recognized for two target audiences, the novice user and the classroom instructor. The QoG is a collection of country-year datasets whose variables include data on population, respect for human rights, and political regime type [20]. Its primary objective "is to address the theoretical and empirical problem of how political institutions of high quality can be created and maintained" [2].

A simple empirical exploration of this question using the QoG data, however, can be fraught with difficulties for novice users. Minimally, one needs to be familiar with some statistical software package and to understand enough statistics to be able to analyze the data. The QoG data would have to be downloaded locally, as would the software being used.

Our tool improves accessibility by reducing or removing these initial barriers. Users interact with the data in visual, gesture-based ways, and so the time spent learning how to use TwoRavens is negligible in comparison to what is necessary to analyze data with R, for example. Statistical models are represented visually as a directed graph, and users may instantaneously view each variable's distribution and summary statistics by hovering over a node in the graph. For example, Figure 3 shows a description of the QoG variable `bl_asy25f` in the left panel. The variables are listed in the left panel, and are added and removed from the modeling space by clicking on the variable name. The modeling space, the center of the figure, shows a representation of a user-defined statistical model.

For instructional purposes, the QoG represents a type of dataset that may be used for a class project. It contains 746 columns of data in a time-series, cross-sectional format. Many of these columns are substantively interesting dependent variables, while others appear as explanatory variables in many models in quantitative Political Science. For instructors to teach how to use data to study questions in politics, requires some degree of expertise in a statistical software package, a projector, and a computer with the rights to that software. For student projects, each student would have to download the data and some statistical software to their personal computer, and the instructor would have to provide materials and guidance on software usage. Most statistical software is proprietary, so students are often restricted to labs that are equipped with the necessary licenses. TwoRavens removes these barriers, and all that is necessary for the instructor to bring data to the classroom, and for students to analyze the data, is an internet connection and a Web browser.

6. CONCLUSIONS

TwoRavens is a Web-based tool for statistical analysis that is lightweight, broadly accessible, and device independent. It integrates with Dataverse, providing access to the tens of thousands of data repositories that exist there, and leverages the power of Zelig, an R library that provides a common call structure for a large number of statistical models. Entirely gesture-driven and intuitive for users of all levels, TwoRavens reduces barriers to statistical analysis and promotes the proliferation of empirical research.

This article details the design of the user interface and its applicability for use by statistical novices and users not familiar with or who do not have access to statistical software. Although integral and foundational to the TwoRavens project, the UI is only the first of three layers. In future research, the TwoRavens project will be adding the model "selector" and results "accumulator" layers to provide more automated guidance on model selection and specification. The selector synthesizes the user input and automates suggestions, such as potential omitted variables, functional form,

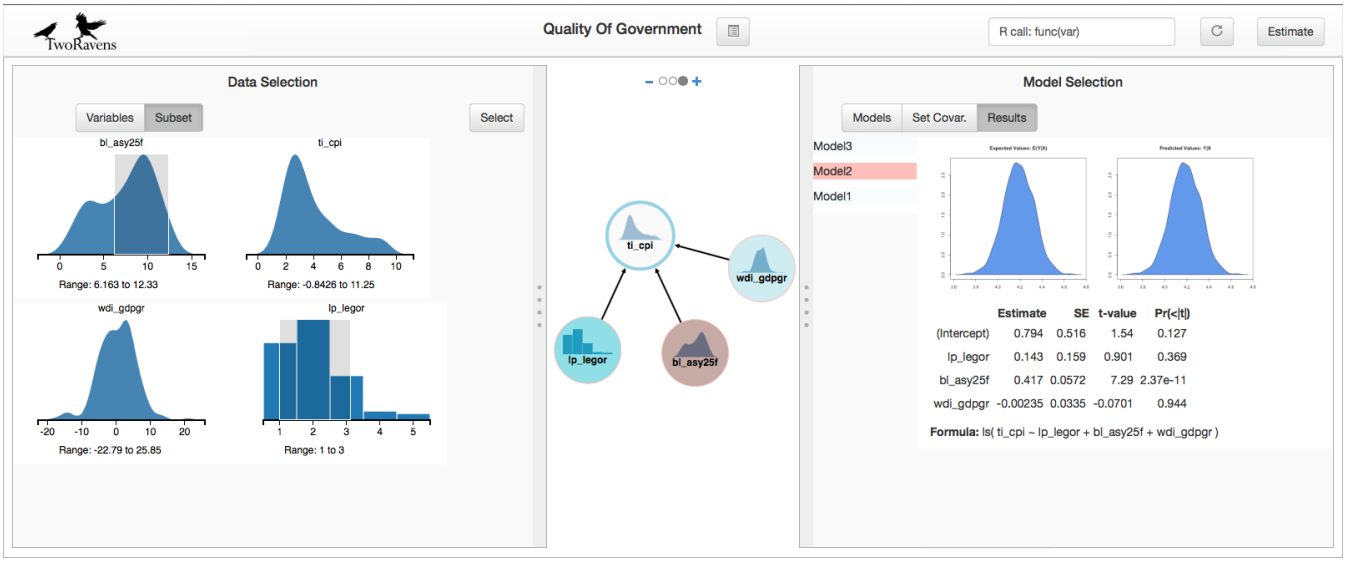


Figure 4: Results Shown in Graphs and Table

and choice of statistical model. The accumulator stores all models that have been estimated on each dataset, and provides users with feedback on existing research using that dataset and datasets judged to be similar.

7. ACKNOWLEDGMENTS

The authors would like to thank Mercè Crosas and Gary King for extensive feedback and ideas, Leonid Andreev, Michael Heppler and Elizabeth Quigley for continued development assistance, and Dwayne Liburd for creating our logo.

8. REFERENCES

[1] Data document initiative. `http://www.ddialliance.org`, Jun 2014.

[2] Quality of Government. `http://www.qog.pol.gu.se/research/`, May 2014.

[3] M. Bostock and J. Heer. Protovis: A graphical toolkit for visualization. *IEEE Trans. Visualization and Comp. Graphics,*, 15(6):1121–1128, 2009.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. Visualization and Comp. Graphics,*, 17(12):2301–2309, 2011.

[5] W. V. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC infectious diseases*, 11(1):37, 2011.

[6] M. Crosas. The dataverse network: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, 17(1-2), 2011. Available at: `http://j.mp/12yqVCZ`.

[7] M. Crosas. A data sharing story. *Journal of eScience Librarianship*, 1(3):173–179, 2013.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

[9] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2009.

[10] D. Edwards. *Introduction to graphical modelling*. Springer, 2000.

[11] K. Imai, G. King, and O. Lau. Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics*, 17(4):892–913, 2008.

[12] G. King. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, 36:173–199, 2007.

[13] G. King. Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics*, 47(1):165–172, 2014.

[14] G. King, K. Imai, and O. Lau. Zelig: Everyone's statistical software, 2007. http://zeligproject.org.

[15] G. King, M. Tomz, and J. Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American journal of political science*, 44(2):347–361, 2000.

[16] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[17] Processing.js. `https://www.processing.org`, Jul 2014.

[18] Shiny. `http://shiny.rstudio.com`, Jul 2014.

[19] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[20] J. Teorell, N. Charron, S. Dahlberg, S. Holmberg, B. Rothstein, P. Sundin, and R. Svensson. The quality of government basic dataset made from the quality of government dataset, 2013. Available at: `http://www.qog.pol.gu.se` [Accessed: 15 May 2014].

Colors of the street: color as an image visualization parameter of Twitter pictures from Brazil's 2013 protests

Johanna I. Honorato
johonorato@labic.net

Lucas O. Cypriano
lucascypriano@gmail.com

Fábio Goveia
fabiogoveia@labic.net

Lia Carreira
liacarreira@labic.net

Labic, Universidade Federal do Espírito Santo
514, Fernando Ferrari Ave. - Vitória, ES – Brazil – CEP: 29.075-910
+55 27 4009-2752

ABSTRACT
This paper aims to discuss color as a methodological tool in the analysis of large quantities of images. For this purpose, this paper presents a series of researches done by two data analysis labs, Software Studies Initiative (EUA) and Labic, the Laboratory of Image and Cyberculture Studies (Brazil), in order to illustrate its different uses. Moreover, this paper shows Labic's recent research on color as a parameter for the analysis of 85.585 images linked to twitter hashtag #vemprarua, an important hashtag related to Brazil's 2013 protests. Thus, this paper highlights the importance of colors as parameters, while identifying issues and contributions to contemporary data science.

Categories and Subject Descriptors
I.4.8 **[Image Processing And Computer Vision]***:* Scene Analysis *– color.*

General Terms
Measurement, Documentation, Design, Standardization.

Keywords
Big Data, Colors, Data visualization, Image, #Vemprarua, Image analysis.

1. INTRODUCTION
The production, dissemination and storage of digital images have achieved large scales with rapid technological advances and accessibility in contemporary society. Image production, with its multiplying variety of tools and available apps for online sharing, has boosted this ever changing scenario, being, therefore, an important and complex contemporary context to be studied and better comprehended.

Differently from contemporary semantic studies (that already is a well developed research field, with its well established tools and softwares), the analysis of large amounts of images is still underexplored, considering that there are fewer tools and researches presently available regarding image datamining, visualization and analysis. Image processing and storing requires great memory capacity and powerful devices, as well as specialized professionals. Although in recent years these processes have become more accessible to all sorts of researchers (with its vast developments and lower prices), extraction and analysis of large amounts of images remains a challenge due to its peculiarities.

In this research scenario, images are analyzed though different data parameters, such as its sharing frequency, time and/or size, in order to create all sorts of visualizations. However, this paper focuses on researches that use different types of color information (such as hue, brightness and saturation) as a parameter for analysis and visualization of image data[1]. Our goal here is to study its importance in revealing a variety of patterns and dissonances that can help us better understand the context and modes of image production today.

Thus, our paper focuses on data collected from the 2013 Brazilian protests #vemprarua hashtag on Twitter, retrieved from the 15th of june to the 15th of July of the same year. The 2013 protests became a large movement that gained the support and participation of millions of people in the whole world. With this large engagement, social media websites gained great relevance, enabling protesters to rapidly share pictures and ideas, and promote a variety of debates and events. It also enabled people at home to become part of this social movement, sharing information and, thus, helping to promote the event and spread the news. Due to its importance to Brazil's social and political context, this paper also aims to better understand its contexts and repercussions though image analysis using color parameters.

2. INTERNACIONAL STUDIES USING COLOR AS A METHOD FOR IMAGE ANALYSIS IN BIG DATA RESEARCH

2.1 Color Analysis in Visual Arts
Visual art, such as paintings, can be one of many spheres in which patterns can be revealed though color analysis. For example, painters make use of a variety of colors to produce their works of art and establish themselves within a specific artistic style. Using color as parameters when creating visualizations of these art works, we can perceive and analyze certain differences between different artists and their works, enabling comparative analysis or even analyze what can be called "stylistic development" of a particular painter.

On this matter, Software Studies Initiative published in June 2011 a research analyzing two visual art collections, one by Piet Mondrian and the other by Mark Rothko. The research was based on their images' visual elements (such as hue, brightness and saturation), thus revealing patterns not only between the works themselves, but also between the artists. The purpose of that particular study was to compare a certain number of Mondrian's paintings to Rothko's produced in similar periods of time in their

Conference'10, Month 1–2, 2010, City, State, Country.

[1] This paper, due to its size limits and piratical purposes, does not aim to present an analysis of the context of visual studies and visual perception, although Labic recognizes its importance to the field.

careers. Through this comparison, the research identified their initial predominant artistic style as being related to the styles of their predecessors. But it also found that, as the years went by, Mondrian and Rothko began to differ their color pallet, which was interpreted by the research as the artists' concerns in developing their own style, therefore, diverging from figurativism[2].

Another image analysis was done by Software Studies Initiative using color in relation to time parameters. That particular study was based on Van Gogh's experience in Paris compared to his time in Arles, and it showed that the set of images from the later contrasted with the set from the former, due to its higher saturation and brightness - a result of the painter's new color experimentations. Thus, the visualization proposed by Software Studies suggests that Van Gogh's paintings were influenced by the spatial changes in his life, as he moved from one city to the other.

2.2 Phototrails' Color Analysis

In July 2013, Nadav Hochman, Lev Manovich and Jay Chow - researchers from Software Studies Initiative -, developed a project called Phototrails. Their goal was to explore, in a planetary scale, visual and dynamic patterns and structures of user-generated contents of Instagram. The study showed, thorough visualizations created using images from that photo sharing online network, how temporal changes and visual features of different locations can reveal their social, cultural and political characteristics, as well as people's habits around the world. In one of their analysis, the researchers chose, among millions of images captured from Instagram, several random samples of various cities, each containing 50,000 images. From that chosen dataset, it was extracted basic visual information (such as average color, brightness, saturation, number of edges, contrast, etc.) to create different visualizations and, thus, highlight each city visual identity in a specific period of time[3].

2.3 Flickr Flow's Color Analysis

Flickr Flow is a project from 2009, developed by data visualization researchers Fernanda Viégas and Martin Watternberg, which serves as an example of image visualization by color using contemporary photographs to retrieve its data. The study started using collections of photographs of Boston Common found and extracted from the photo social network Flickr. With their available data, the researchers divided all photos by month and calculated their colors' relative proportions. The projects following step was to, then, plot a "wheel" shaped dataviz using both color and time as its parameters[4].

Thus, as a result of the visualization created, differences between the seasons of that particular year can be identified through its color variation pattern. At the bottom of this visualization, it's possible to identify a great amount of grays, whites and lighter colors, which represent winter. One can then observe, clockwise, the increase of more vivid colors (variations of pink, purple, green and yellow), thus representing spring. Following this pattern, one can also observe the other seasons, with fall being indicated by the yellows and oranges and summer by large amounts of bright colors and a very few of white tones.

[2] Images and more infos on the study are available at http://lab.softwarestudies.com/2011/06/mondrian-vs-rothko-footprints-and.html

[3] The research results and images are available at http://firstmonday.org/ojs/index.php/fm/article/view/4711/3698

[4] The research results and images are available at http://hint.fm/projects/flickr/

3. #VEMPRARUA's COLOR ANALYSIS

3.1 2013 Brazilian Protests and its hashtag #vemprarua

The previous studies conducted by Software Studies Initiatives has shown the vast possibilities regarding image analysis using colors as parameters, bringing great contributions to data science. Taking in consideration the contribution that they have also made to the analysis of social and cultural behavior and patterns through image visualization, Labic has been developing, using other tools and visualizations, a research in which we can better understand the complexity and variety of political and social issues implicated in the emergence of June's 2013 protests.

The objective of this study, named "Visagem", is then to analyze Twitter hashtag #vemprarua (which can be translated as "come to the streets"), an iconic expression of the Brazilian protests and, thus, the most used hashtag to refer to this particular social movements within social media websites. The 2013 protests in general were against government corruption, poor government financial administration associated with 2014 World Cup, and also for better quality of transport, security, education throughout the country. It is then an important social and political movement that deserves especial attention and research.

3.2 #Vemprarua's Datamining Process

Our datamining method was based on the retrieval of data from the popular social networking service Twitter through a software called yourTwapperKeeper (a.k.a YTK), which uses Twitter API to gain access and extract the necessary data. With this method, all tweets that had the matching hashtag *#vemprarua* were collected, creating a csv file with all the available information (such as who tweeted, date of publication, number of retweets, etc.).

With this csv file, Labic used a Java based script called *Crawler*, developed by our lab and whose function is to separate tweets that contain links from those that don't. After this process, the script access each tweeted link and captures the images that obeys the parameters set previously by our researchers, such as a minimum size of 15 kB or 200 x 200 px, and the extension files PNG, JPG, JPEG, TIF or TIFF.

Between June 15 and July 15 of 2013 (a critical period in that year's political and social protests), we extracted 85,595 images, originated from a total of 404,006 tweets. These images, despite only being retrieved from Twitter, came originally from a variety of websites and apps, such as online news websites, blogs, and other social networking websites, that was then shared by several social media profiles.

3.3 #Vemprarua's Color Parameters

In order to analyze this large amount of images, HSB color scale was used in this research, in which the color of each pixel in a image is composed by three numeric data: hue, brightness and saturation. Basically, hue values goes from 0 to 255 (equivalent to 0º to 360º degrees), thus forming a color circle. Brightness is then determined by values ranging from 0 to 100, in which zero means no light (black) and 100 means maximum presence of light (white). Finally, saturation follows the numerical variation of brightness, also ranging from 0 to 100, however, being 0 an absence of tone (presence of grays) and 100 being fully saturated colors (no grays). This chosen color scale basically helps identify groups of images with close measurements, and also allows image organization using these same parameters.

With this issue settled, the plug-in "Measure" (a plug-in of the software ImageJ) was used in order to read the values of each pixel and then calculate its hue, brightness and saturation. Thus, it

was through these three color parameters that this study was able to develop different visualizations and analysis of large amounts of images from the #vemprarua movement, enabling the researchers to identify certain patterns and characteristics.

3.4 Analysis of #vemprarua 's Image Visualizations

With the images captured through YTK and with the visual information gathered by the Measure plugin, it was then possible to plot different visualizations in which these large volumes of data can be compared. In order to make these plots, we used a software called ImagerPlot, which was developed by Software Studies Initiatives, housed within the UCSD Division of the California Institute for Telecommunication and Information Technology.

3.4.1 Brightness x Saturation

With these plots, which visually highlights sets of images separated by its color parameters, an analysis was made possible. In the visualization below (Figure 1), #vemprarua's images are distributed throughout three major groups: a whiter set, mostly found in the upper left quadrant; a darker set, found at the base of the this dataviz; and a more colorful set, on the right upper quadrant.

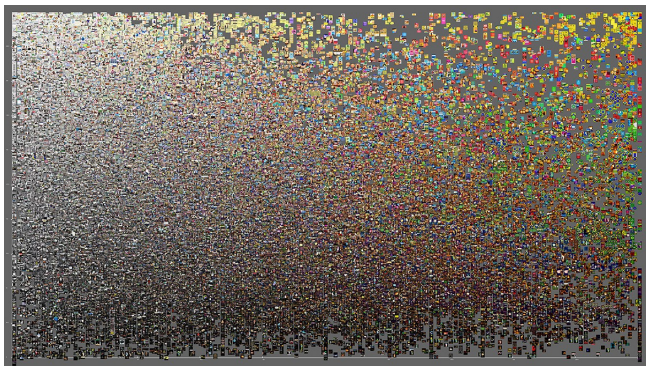


Figure 1 - 85.595 images sorted by X-axis (saturation median) and Y-axis (brightness median)[5]

In the first group with predominantly white images, a greater presence of posters, prints of documents and newspapers covers that have been shared throughout the months in which the protests occurred can be noticed. The distribution and circulation of information of this type was predominant through June 15 to July 15, where, among other contents, it is possible to find: posters aiming to motivate people's participation in the protests, as well as to sharing of more information on the objectives and schedules of the events; documents and newspaper covers that contained information from mass media; and others.

The second group consists in images taken during the protests. They are images of a grayish tone, due to the predominance of the streets' asphalt color, visible during the daytime as well as at night (however, with a darker tone), thus beeing one of the most striking features of these events.

The third group is what aggregates posters and advertisements attached to the contents shared with the #vemprarua hashtag. The posters in this group moves away from the previous black and white pattern, towards more vivid colors: mostly blue, green and yellow, thus largely associated to Brazil's national flag.

3.4.2 Hue x Brigtness and Hue x Saturation

The next visualizations (Figure 2 and 3) arranged the images in #vemprarua's dataset according to its color bands when the parameters were modified to "Hue" (X axis) and "saturation" (Y axis). Thus, groups of similar images are clearly marked and the appearance frequency of certain types of images throughout the collection are better understood. The tracks that stands out are red, orange, yellow, green, blue, and the combination of purple and pink.

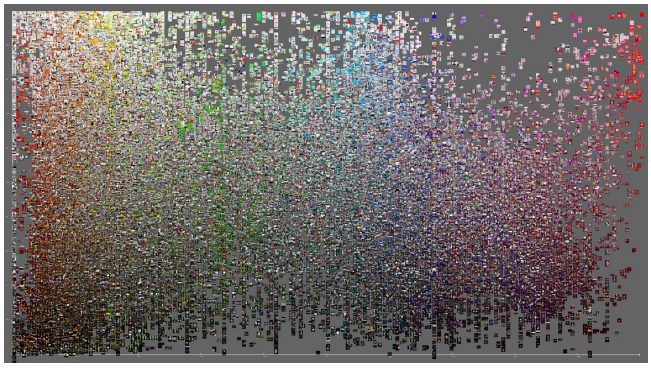


Figure 2 - 85.595 images sorted by X-axis (hue median) and Y-axis (brightness median)[6]

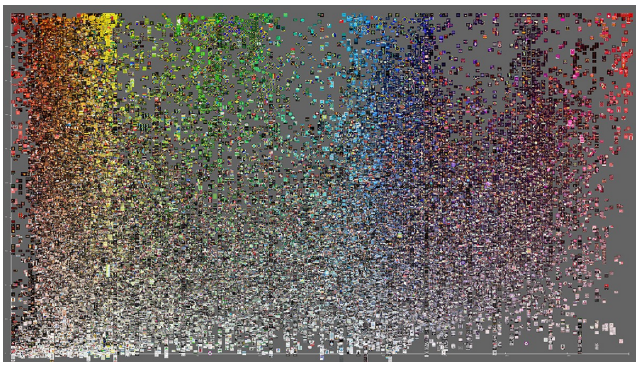


Figure 3 - 85.595 images sorted by X-axis (hue median) and Y-axis (saturation median)[7]

In these visualizations, the first color range has a larger number of images if compared with the rest of the dataset. The predominant images that appears in this group are photos taken at the time of the protests, even if they were shared later on by the users. With a closer look to the predominant orange tone area, images that are characterized by the street's yellow-orange lighting can be observed, as well as photos of confrontation between police force and protesters, which often involved fires being set, explosions and rubber bullets fired by police and captured by the lenses of the vigilant photographers and protesters.

The green color range is basically formed by the reproduction of the national flag of Brazil and also compose by its re-appropriations: these images varies in size, color and type, and occasionally inserted into green colored posters. On some of these posters, the white band that bears the inscription "Order and Progress" (Ordem e Progresso) in Brazil's flag was replaced by,

[5] It's important to notice that the visualizations here presented are created to be visualized in a larger digital visual devices that enables zooming features and user interaction. For a better visualization experience, this image is available in high definition at http://zoom.it/G8jg

[6] Available in high definition at http://zoom.it/QCYi

[7] Available in high definition at http://zoom.it/tAaW

"In Progress" (Em Progresso), meaning that the country was in a state of change led by the people.

The blue color range is also mostly composed by images of flags of Brazil, focusing on its inner circle. This group also has a lot of photos from Instagram, due to one of its available filters. The last color range covering the pink and purple tones are pictures of the protests that were intentionally faded (with the use of filters, for example) and posters intending to represent a more feminine approach.

3.4.3 Color Visualization by Hue with Static Brightness and Saturation

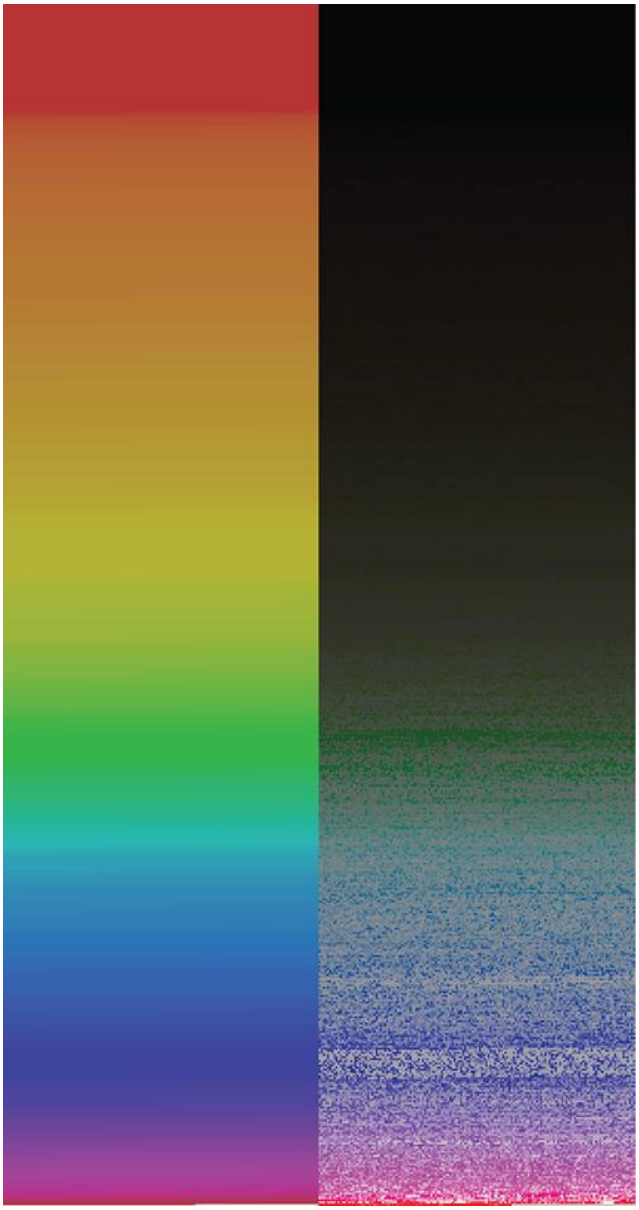


Figure 4a and 4b - Visualization using the median values of hue, saturation and brightness

Instead of placing an image in a specific position determined by its color parameters, we proposed in Figure 4 two different kinds of visualizations. Therefore, for these visualizations we used two types of sets of color parameters. For the first dataviz, we used the hue median, saturation median and brightness median, as to compose the color of the squares representing each image. For the second dataviz, we used just the hue median of each image; the saturation and brightness were established through a standard value. So in Figure 4a we have a visualization of all color parameters of each images, and in Figure 4b it's possible to see more clearly just the hue value. These visualization were created using a script developed in our lab through Processing, in order to visualize the colors medians previously calculated by ImageJ. In these dataviz, each image is thus represented by a square of 2 by 2 pixels, in which the top represents the images with hue median 0 and the bottom, images with hue median 255. Thus, as a result, these recent visualization developed at Labic highlights the large color variations of the Brazilians 2013 social political protests, showing its characteristic visual aspect.

4. CONCLUSION

Throughout this paper, we aimed to understand the importance of the usage of color as parameters for visualizing large amounts of images. Both in the artistic field and in the studies of social movements, color value can reveal more than numeric information: it can also highlight their characteristic visual aspects or styles, as well as point out patterns and singularities of their datasets. As shown in this paper, they can be viewed singularly or in comparison to other images sets. In both cases, color parameters presents themselves as a relevant and simple method for big data analysis.

However, visualizations made with softwares such as ImageJ have certain limitations that can hinder a deeper analysis of the datasets. Using this kind of visualization tools, image plots can only be made using only two coordinates (X and Y). Thus, when a image has the same coordinates as another, an overlapping occurs and you lose, therefore, visual information. Considering this problem, we perceive a need to create a tool capable of adding a Z axis allowing a 3D environment, where image information would not be lost and user interaction is enhanced.

On the other hand, when using Processing, each image is represented by its corresponding pixel, and thus no overlap occurs. In the dataviz presented in 3.4.3 (Figure 4a), each color range in the dataset is clearly represented, confirming the theory that previously observed through the analysis of the first visualizations created with ImageJ (Figure1, 2 and 3), that mostly orange toned pictures were shared during the protests. This fact emphasizes the frequent need of different visualizations of the same dataset for comparison in order to identify or confirm certain characteristics.

The second dataviz in topic 3.4.3 (Figure 4b) follows the same principles of the first, in which the image is represented by its pixels' color, brightness and saturation values(HSB), showing that despite the predominant color being orange, the protests' general tone is dark, which again refers to specific characteristics of the June protests: being an predominantly evening event.

Thus, this paper acknowledges that visual characteristics (such as hue, brightness and saturation), when used as a parameter to organize large amounts of images, can reveal artistic patterns and also social, cultural and behavioral patterns. Therefore, image visualizations using color parameters can present more then numerical values, also pointing to various perspectives of an determined event of practice.

5. ACKNOWLEDGMENTS

Our thanks to the Federal University of Espírito Santo (UFES), the National Council for Scientific and Technological

Development (CNPq) and the Espírito Santo Research Support Fundation (FAPES) for the this research financial support. This research is part of the "Visagem" project of the Laboratory of Image and Cyberculture Studies (Labic).

6. REFERENCES

[1] Hochman, N., Manovich, L., Chow, J. 2014. *Phototrails*. Available at: < http://phototrails.net/>.

[2] Manovich, L. 2012. *Data Visualization and Computational Art History*. Available at: <http://lab.softwarestudies.com/2012/04/data-visualization-and-computational.html>

[3] MANOVICH, Lev. 2011. Style Space: *How to compare image sets and follow their evolution*. Available at: <http://lab.softwarestudies.com/2011/08/style-space-how-to-compare-image-sets.html>

[4] Manovich, L., Hochman, N. 2013. *Zooming into an Instagram City: Reading the local through social media.* First Monday, v.18, n7, 2013. Chicago. Available at <http://firstmonday.org/ojs/index.php/fm/article/view/4711/3698>

The use of modularity algorithms as part of the conceptualization of the perspectival form in large networks

Lorena Regattieri
Labic - UFES
Federal University of Espirito Santo
regattie@ualberta.ca

Jean Maicon Medeiros
Labic - UFES
Federal University of Espirito Santo
jeanmrmedeiros@gmail.com

Fabio Malini
Labic - UFES
Federal University of Espirito Santo
fabiomalini@gmail.com

ABSTRACT

How can we identify perspectives in large networks through the application of modularity algorithms? In the digital humanities [1][2], there is a fair number of scholarly work exploring computational routines to cluster and analyze enormous amounts of data. Recently, social data became a valuable source to study collective phenomenon, they provide the means to comprehend human collectivity by using graph network analysis. In this paper, we describe our approach on the manner of post-social anthropology [3] and social sciences using technical methods: quantitative analysis and modularity optimization. The computational turn is part of the ongoing process to conceptualize the "perspectival form", as the other would be the semantic analysis of the qualitative data. This technique uses a python script to extract the co-occurrence hashtags network from a Twitter dataset in order to apply in the context of the open-source software Gephi. Our experiments successfully exhibit how social networks can be unfolded when submitting a sample dataset of hashtags to the procedure found in the critical dimension of computational models. Therefore, it discovers the flow of perspectives when the strategy is follow in new workspaces, creating then categories that reveals points of view underneath the controversy. Concluding, this study presents a theoretical and methodological framework based in the post-structuralists, a composition that aims to support studies in different fields of social sciences and humanities.

Categories and Subject Descriptors

D.3.2 [**Programming Languages**]: Language Constructs and Features – *abstract data types, polymorphism, control structures.*
I.5.3 [**Pattern Recognition**]: Clustering - algorithms, similarity measures.
J.4 [**Computer Applications**] Social and Behavioral Sciences – Sociology

General Terms

Documentation, Human Factors, Theory, Algorithms and Design.

Keywords

Post-Social Anthropology, Network Science, Amerindian Perspective, Modularity Algorithms, Complex Networks.

1. INTRODUCTION

This paper understands that social networking is an anthropological phenomenon. A graph of social networking is a material representation of human relationships. Therefore, both the algorithm that seeks to analyze them, as the natural language vocalized on them, are in continuous process of interrelation to interpret the social world. The algorithm alone does not explain these relationships. But collective action, today generative of digital traces [4] cannot be explained alone, only with historical social theories of the humanities.

Graph clustering or community detection [5][6][7][8] in complex networks have a long history of research in machine learning and graph theory [9]. The studies in the field have gain attention from several areas, the most common studies are find in biology, technological, and physics. In the meantime, the literature in Natural Language Processing [10][11] and Probabilistic Neural Networks [12] have shown us the possibilities in document modeling, text classification, and collaborative filtering for large corpora.

In this paper, we describe a certain method developed by researchers at Laboratory of Studies in Images and Cyberculture (LABIC)[1], located at Federal University of Espirito Santo (UFES), Brazil. It consists in being a simple, but efficient and peculiar method developed to support studies in social sciences and humanities. Our novel perspectival framework uses a Twitter dataset publicly available online, thus, a variety of 500k+ tweet twitter feeds are draw on for examples. Such method uses Gephi [13] and its algorithms, resulting in visualizations and statistics. The method aims to find communities on a network formed by co-occurrence of hashtags in a tweet, in other words, we set a network of hashtags in order to compose a multiplicity.

The relevance in the contemporary context of online network sites serves as the means to interpret the political and collective actions, that is why Twitter is our "field" of work. We consider the social network a rich terrain of dispute, noticing the many uprisings around the world: #OccupyWallStreet, #15M, #OccupyGezy, #VemPraRua, and #NãoVaiTerCopa. Other social phenomena can be considered a perspective in progress, like #ClimateChange. While recently proposed methods practice detecting topics in historical and literature corpus by using probabilistic topic modeling [14], we aimed to present a new methodology to underline not just a topic model procedure for digital data, but to reveal the points of view in constant flow, in fact, profiles in a battlefield.

In order to comprehend the layers of texts in the digital traces left by humans, we rely in the actor-network-theory [15]. The main idea is to work in the same level of both, the actors and its

[1] http://www.labic.net

attributes. “A network is fully defined by its actors." [16] ANT and network analysis provide the argument to study digital data without worrying about the standpoint of the individual or collective. It is possible to negotiate to one level to another, from the parts to its whole, only by continuously rearranging the actors, or the nodes. There is no overlapping, it is matter of reorganizing ones positioning. The cartography of controversies [17] is the didactical application of the ANT, it serves as a range of techniques to explore public debates. Observation and description is essential to the scholarly work done in this paper. In this meeting between computing methods and the post-social anthropology [3], the Lautorian socio-technical networks approach will support the process of revealing points of view in disputes.

Our methodological framework poaches the Amerindian Perspectivism [18] to find the foundation for our ongoing experiments to compose a "perspectival form" in large networks. Again, they are called large networks because they are made of thousands or even millions of nodes and edges. Most importantly, comprehending the node as a social profile in the network, thus, the edges, as the link between One and the Others. Then, a network is only constituted by the existence of the other. Eduardo Viveiros de Castro subverts the idea we have of cannibalism, which is an idea that guided in the conception of "to cannibalize" the other is to eat the other. He inverts the enunciation, saying that cannibalism is a way out of self to go into the other, for each other. The node as a profile on the social network it increasingly comes out of the self to "retweet" what is better or worse from another, therefore, assuming the point of view of that other (and they are of many types). Nowadays, the other is the element that captures us. It is an anthropological turn, which we live in.

In fact, this is our inspiration to reconceive a qualitative-quantitative method of analyses throughout machine steps, which we know in computing as the algorithm. When applying these procedures to comprehend collective phenomena, it produces new perspectives and methods. The computer requires the cascade of texts and hashtags we collected in our dataset to metamorphose into the grid of numbers. [19] The framework we have been testing is based in the Louvain algorithm [20], in which we compute to maximize the network modularity.

The use of Twitter, in particular, has led us to a couple of challenges in text clusterization process. As the qualitative research process evolve and the number of tweets increases to millions, categorization and the topology of the network became a problem. “The whole is always smaller than its parts”.[16] A large network features an illusory representation. It overlaps itself in distinct layers, social groups and thoughts, as if was part of a single network topology. In theory, the social is crossed by a multiplicity of natures, perspectives, worldviews, produced by different human groups. And here is our hypothesis: thereby, every network is, rather, a network of perspectives, which are usually in dispute.

The methodology that first was based in data mining and clustering thousands of words needed a new framework. Given this problem, we created the hashtag network script. After the consultation of literature available [21] new possibilities have rise, from the initial goal to find a method to fastening clusterization of words and categories to the use of hashtags to find perspectival forms. Nowadays, the discussions indexed to a hashtag often become themes of conversations between halls. The hashtag, based in our tests, prove to be the better solution for social scientists working with data science. When using the hashtag sign, the user is segmenting a topic of interest, more than that: he allies itself to a point of view on a subject. It is simple to analyze that once someone have generated a tweet and already used a hashtag, it is as if the user is already categorizing the text for the researcher. In addition, the hashtag represents the existence of a debate that matter or even just some cause that people aimed to call attention for it. Either way, the many ways that people give meaning to points of view by indexing value to a specific word will qualified a perspective in the public debate.

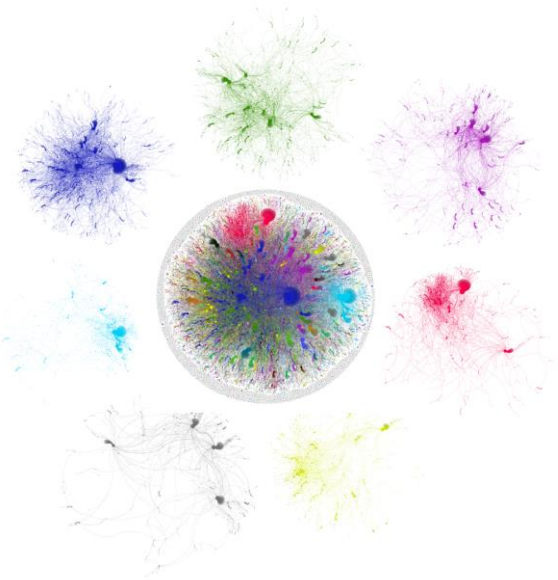


Figure 1: The figure shows the center of the network #VemPraRua, consisting of 125 000 Retweets. Only when analyzing the perspectives (networks around the center) it is possible to understand the different perspectives on the network.

2. THE ANTHROPOLOGICAL THOUGHT AND NETSCIENCE

The substance of our framework is in how we interpret modules without changing the levels or scale of plan. In online social networks, we argue the existence of movements and circulation in a flat surface with no consideration to hierarchy. The node is situated in the terrain of dispute, one that is only defined by its network.[16] In this case, when exploring the dots in the graph, which in our dataset are the hashtags, the actor moves to the network, interacting with others in the same level. This is where we stand with Latour, in a flat ontology.

The approach we reclaim to study online networks is the one inherit from Pierre Clastres.[3] In any case, we propose a descriptive study of a terrain which we understand to be in constant dispute. This allows us to rely once again in the indigenous world, which there is a surviving violence itself, a reference to problematize the thesis of repulsion and attraction of the algorithm of modularity. In short, we make use of the concept of cannibalism, which derives from the complex notion of cannibalism. Applied in the field of hashtags as views, this very cannibalism lives of the perspectival forms within the network

revealing then a mode of operationalization. This is a process of maximal reduction of one single node and another, almost like a microscopic work to see the minor points of view. "Exchange, or, the circulation of perspectives: exchange of exchange, that is, change." [22]

In data science, complex networks [23] are identified as very large networks, millions or billions of nodes and edges. This sort of networks occur in different contexts, it is possible to recognize in nature, society, technology, economics, etc. One of its fundamental characteristics is the temporal evolution aspect. Complex systems constitute themselves of many non-identical elements connected by a diversity of interactions. Several networks in nature, ecology, economics, human relationships in social networks and the web has the same topological structure. They are known scale-free networks [24]. We will associate this computational concept with the understanding of networks from Bruno Latour.

In this sense, the actor-network theory (ANT) comes in hand with the inquiry we propose. The large networks in this empirical study come from the NET, which we purposely stress in the same way Latour does with ANT. To trace the circulation and interactions of points of view and objects, ANT is going to explore the constitutive connections between actors (the actants), both animate and inanimate, and the generative potential of those interactions. In his own words, "(…) network does not designate a thing out there that would have roughly the shape of interconnected points, much like a telephone, a freeway, or a sewage 'network'… It qualifies its objectivity, that is, the ability of each actor to make other actors engage in unexpected relations."[15] More precisely, we consider social profiles as living things. Often happens that in the information networks, it is not possible to recognize the "form", only the information. By that we meant the profiles that uses the language like a human component, but notice, they are only information, or robots to act as man. However, the meaning arises from the disparate actions. [27]

We mend our theoretical foundations in the connections we perceive between anthropology and post-structuralism. Which summing up is circumscribed in the post-social-anthropological net of authors listed here, considering then the deleuzian concept that comes from the mathematics, where we find the means to comprehend the multiplicity as a point of view. It creates a new kind of entity, rejecting any generalizations, the one we know as 'rhizome'. Therefore, a rhizomatic multiplicity does not, in fact, behave as one, because it is not possible to do that when it operates as assemblages of becomings. Here is when Latour meets Deleuze and the notion of actor-network, one which the network cannot be one thing, yet, again, because anything can be considered a network.[22] And finally, in the next section, building up from this interdisciplinary dialogue, we present how the amerindian perspectivism support our hypothesis in exploring the complex world of large networks, finding a perspectival form within the modularity algorithm.

3. THE PERSPECTIVAL FORM WITHIN THE MODULARITY

We were called into the indigenous world to reflect the network studies, mainly due to a natural notion of multiplicity in the indigenous society.[26] Primarily because we have for long studied in information networks, a political aspect that we find in the modes of existence peculiar to the indigenous society, a way of existence, i.e., a substantially minor of existence, in a minority character. Therefore, we are concern with the mechanisms that inhibit or block the emergence of a totalizing discourse. Therefore, "perspectivism does not state the existence of a multiplicity of points of view, but the existence of the point of view as a multiplicity." [27]

Modularity is one of the possible measure for detecting communities in complex networks. A set of nodes categorize itself as community by its modularity if the fraction of links between them is higher that expected ia network called "null model", which is used as a reference. [28]. A complex network with a high modularity indicates strong community structure, in other words, the nodes inside the same community has a dense connectedness and has a sparse connexion between other communities.

The algorithm applied in this paper to find communities, since we use Gephi [13], is the Louvain Method. Such method does community detection in weighted graphs and has characteristics such as greedy heuristic, local optimization of modularity, very fast (complexity O(nlog(n), n: number of nodes), non-deterministic, return hierarchical partition. The Louvain Method is an "algorithm that finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection." [20]. Think of the network as a perspective. Well then, the nodes that compose such network will form an alliance, ie, they will form a covenant relationship between viewpoints. The link between two nodes is exactly the distance between them, and also, the distance between points of view. It turns out, then, that the way which we apply the algorithm maximizing the modularity, the network is partitioned into modules, testing all nodes until no node can belong to another module. It is a dimension of alterity, the same as found in Amerindian perspectivism. "Perspectives encourage you to believe OUT of them." (Roy Wagner)[2] The algorithm repeats this process of exchange and change, successive times for all nodes. Autophagy is a survival of hashtags in the network. A roundup of alliances.

4. METHODOLOGY

"The object as such: why a perspective is not a representation"[31].

The first step of the method is, of course, to have the dataset to be analyzed, the collection of tweets formated in a comma separated file (csv). The tool utilized to get these tweets is called yourTwapperKeeper[2]. The procedure begins with the choice of a term or hashtag, the tool does the job of archiving the massive amounts data. This process provides a historiography of what have been vocalized related to the research expression. With enough data to go through ethnographic rendering, we can go to the "field", which for us means to explore a database of entities and attributes.

The second step is data processing. As we know, hashtags are one of the most commonly used form of categorization and indexation among users in social networks, such as Twitter and Facebook.

[2] http://www.github.com/540co/yourtwapperkeeper

One can say that the hashtag summarize the content of the tweet, positively or negatively, confirming it or contradicting it. So, this next step consists in creating a "Hashtag network" from the tweets previously collected. The Hashtag network is a complex network that links hashtags if there is co-occurrence between them in the same tweet and it forms a weighted network, as it can happen twice with the same hashtags. The creation of this complex network is provided by a script programmed in our lab and its output is a csv file that will be used in the data mining process.

The third step relies on drawing the network and manipulating with its structure. In order to visualize the network, we import it to Gephi. For now, the first view of the network is a hairball, a completely unintelligible graph. This is the time when modularity comes into the picture. But before that, there's a very important act. We will have to delete the "main node", in other words, the hashtag that links all nodes. Therefore, the next move is to apply "Modularity", set the parameters of your choice and wait until calculation is over. Next step, applying the modularity class calculated for each node and thus forming the communities. One way to apply it on the network is setting the colours to the nodes, thereby emphasizing the communities, in our case, the topics of discussion. The next important move is to calculate the "Average Weighted Degree" which gives the user a way to apply different sizes to the nodes from their weighted degree, and this was the next step. The network isn't longer a hairball and the recognition of communities is clearer, thus, as for the biggest nodes in each community, they define the points of view of that community.

Lastly, each community is a network of point of views and they are distributed through Gephi's workspaces. Now, we apply the modularity and calculate the average weighted degree again. The final touch consists in setting the design of the graph with the "Circular Layout" option, it is also more visually interesting to order the nodes based in the modularity class. We advise for matter of design to find the node with higher degree, in which we will identify the most prominent point of view of the particular network. By now, we expect for terms of visualization and exploration to have a network of hashtags, i.e, the perspectival form of the network.

4.1 The case with the #WorldCup

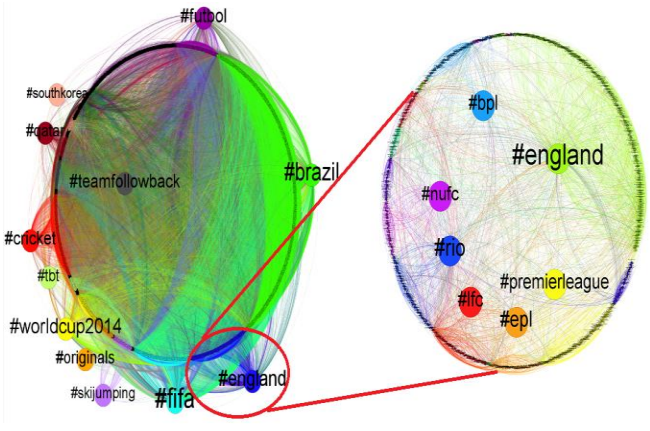


Figure 2: #worldcup's main perspectives and #england perspectives on #worldcup.

The dataset consists in 271.013 tweets that were collected between february 4th and may 4th, 2014. This image is a view between acts in the third step of our method, after the first pass of the modularity optimization algorithm and rearrangement of the nodes with highest weighted degrees in each perspective. It is an overview of #worldcup's hashtags network as the main perspectives are emphasized. As we can see a certain noise or distortion is identified in the network, as in "#cricket", where the hashtags mean to mention the cricket world cup, or in #teamfollowback, where users tend to flood their timeline in order to get more followers.

In this perspective of the network (Figure 2), it is visible the english topic being discussed. The different subtopics, evident among the nodes, make this assumption clear. And so, as seen in the hashtags #epl, #bpl and #premierleague, meaning the discussion of the English Premier League a.k.a. the english national championship, and in #nufc and #lfc, meaning Newcastle, United FC and Liverpool FC, both english teams, and last, but obviously not least, the hashtag #rio, that clearly connects the main discussion #worldcup, as the English team is going to train in the Rio De Janeiro city before the cup.

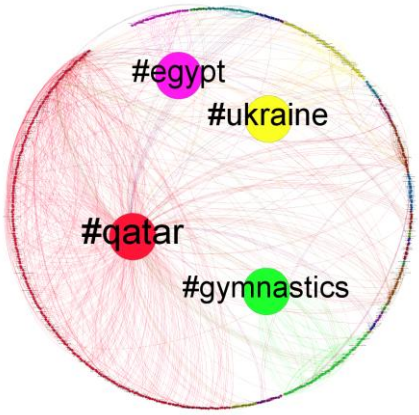


Figure 3: #qatar perspectives on #worldcup.

After emphasizing the nodes with highest weighted degrees, the human interaction, as research, is truly required to engage the process of perspective perception. The hashtag #ukraine involves the perspective of protests and their recent history with russia, the multiples hashtags are seen in the composition of point of views.

We can identified the following words: #crimea, #sanctions, #russiainvadesukraine, and #worldwar3. But also in this perspective, there is fractal element, because we can also foresee the hashtags #wc2018, #2018worldcup, and #worldcup2018, which suggests that people are already expressing concerns on the country that will host the next world cup, in 2018. As for #gymnastics, the perspective lies in the gymnastics world cup that happened in doha in 2014, which can be seen as noise in our main investigation. And in #qatar, where the 2022's world cup will be hosted, the multiplicity, as point of view, is focusing on several discussions involving #humanrights, #workersrights, #slavery, and such.

4.2 The case with the #ClimateChange

The dataset on climate change was collected between February, 2nd and May, 5th of 2014. In total, we have exactly 1.048.576

million tweets. To analyze the data, we put together a hashtag network of 21.415 nodes.

The number for the hashtags provides a sample of the "heat" of the debate online. In the Figure 4, we had only computed the modularity the first time, the graph display the partition of the network into modules. The points of view with higher average weighted degree indicates as results: #carbonbubble, #energy, #obama, #tcot, #nsa, #gree#, #news, #ows, #truth, #obama, #bbcnews, #fracking, #travel, #jobs, #earthday, #organic, #climate and #climate2014. Who is what in this network? Appearances can be deceptive, although, a few interesting revelations appears already. For instance, #tcot means Top Conservatives on Twitter, this network has a longer effect in the network because it has has an alliance to american Tea Party.

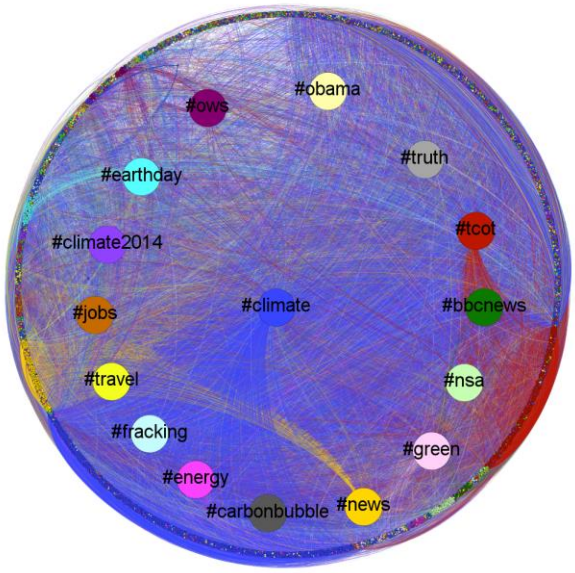


Figure 4: #climatechange perspectives.

Still, note that we have design the perspectival forms in order to visually demonstrate the capacity of some point of views to establish more regimes of alliances. In this orange network of point of views, the high value of internal modularity, clearly echoing the american Republican Party tongue. At the same time, the green network maintain a link to the orange network, the multiple points of view embedded in this green network are #globalwarming and #deniers. No wonder, this perspectival form preserve this alliance with American conservative party.

The blue network proposes a perspectival form of the anthropocene. A hahstag itself, #anthropocene reflects the currently reality of concerns brought by the notion of Gaia. Bringing issues like # energy, # food, # weather, a dimension of the ecological crisis. The reflection of man before the outburst of Gaia. In this case, the blue network has links to the different perspectival forms, such as the #cdnpoli, a network of the point of views involving the environmental crises in Canada. In there, we can find the #KXL #KeystoneXL, the hashtags used about the oil debate.

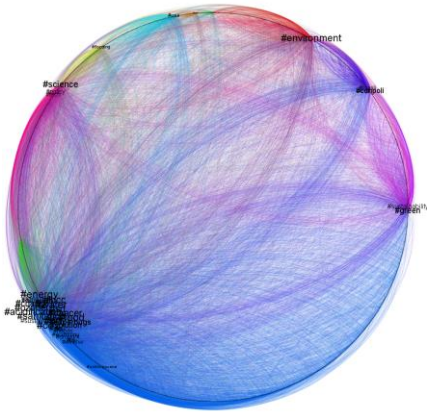


Figure 5: The blue network arises as a perspectival form with high modularity.

5. CONCLUSION

In this paper we have presented theoretical references in Post-Social Anthropology and Complex Networks to support our methodological framework for studies of social information data. Twitter is a rich field of productions, it can create alarming discussions over the necessity to debate the ecological crises, such as the hashtag #climatechange. There is a social memory within the hashtag, that's why in this research we addressed the exploration of points of view though the hashtags in the network. However, the hashtag is also a fictional character that brings together a collective memory and puts it to act in the public space, influencing the understanding of what we understand to be reality. This is not a simulacro 2.0, it is a practice that activates a mode of human existence, the fictional, to expand our critical capacity.

In the case of #climatechange, we confirmed the existence of a variety of networks in the large network. Different perspectives that are completely distinguishable. Such as, the distance between #actonclimate and #teaparty.The analysis of the #worldcup assemble the perspectival form as a multiplicity. Inviting us to dig into the point of view, emphasizing that it is not possible to generalize the network. This procedure, that analyzes the co-occurrence of hashtags in a dataset of tweets, leaves behind tweets with no hashtags and one hashtag only. This implicates on a certain limitation for the method, but also it focuses on its main goal: to study the connection between the hashtags of a tweet and perceive the perspectival form originated by its connections on a complex network.

We describe the intercorrelation of algorithms and the humanities, together it composing a powerful tool that allows a routine of data mining, processing, and visualization of social information. Applying our research methodology has evidenced our hypothesis since it indicates that there are variety of points of view, so a more detailed study of network demands to take into account the perspectives of the network. It is also important to note, perspectives converge in the same direction, so the groups are well defined in which side it defends. Our method indicates that research involving informational networks, such as studies concerning degree, sentiment, hub and authority, which do not take into account the perspectives in dispute in the networks, will tend always to reach conclusions that privilege the richest nodes with more connections. For future work, we plan to refine our

methodological frame with tests in other datasets and to improve the visualization of the perspectival form of the network.

6. ACKNOWLEDGMENTS

Funding for the project generously supplied by National Council for Scientific and Technological Development (CNPq), National Academic Cooperation Program (Procad), Coordination of Improvement of Higher Education Personnel (Capes), Foundation of the Ministry of Education (MEC). Our thanks to the team at the Laboratory of Studies in Image and Cyberculture (LABIC) for the ongoing support.

7. REFERENCES

[1] Moretti, F. 2013. *Distant Reading*. London: Verso. 254 pp.

[2] Jockers, M. 2013. Macroanalysis: Digital Methods and Literary History. University of Illinois Press. 208 pp.

[3] Viveiros de Castro, E, Goldman, M. 2012. Introduction to Post-Social Anthropology. In *HAU: Journal of Ethnographic Theory* 2 (1): 421-433.

[4] Latour, B. 2007. Beware your imagination leaves digital traces. In Times Higher Literary Supplement, 6th April 2007.

[5] Lee, C, and Cunningham, P. 2013. Community detection: effective on large social networks. In *Journal of Complex Networks* (2014) 2, 19-37.

[6] Elhadi, H, and Agam, G. 2013. Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods. In *SNA KDD 2013 International Workshop on Social Network Mining and Analysis held in conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 2013.

[7] De Meo, P, Ferrara, E, Fiumara, G, and Provetti, A. 2011 Generalized louvain method for community detection in large networks. In *Intelligent Systems Design and Applications (ISDA)* 88-93

[8] Fortunato, S, and Barthélemy, M. 2006. Resolution limit in community detection. In *Proceedings of the National Academy of Sciences of the United States of America*. v 104, 1, 36-41.

[9] Milkov, E, Cohen, W, and Ng, A. 2006. Contextual Search and Name Disambiguation in Email using Graphs. In *SIGIR*.

[10] Chang, J, Boyd-Graber, J, and Blei, D. 2009. Connections between the Lines: Augmenting Social Networks with Text. In *Refereed Conference on Knowledge Discovery and Data Mining*, 2009

[11] Blei, D. 2014. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application* 1:203-232, 2014.

[12] Ciarelli, P, Oliveira, E, and Salles, E. Multi-label incremental learning applied to web page categorization. *Neural Computing and Applications* 24(6): 1403-1419 (2014)

[13] Bastian M., Heymann S., Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. 2009.

[14] Mimno, D, and McCallum, A. 2007. Mining a digital library for influential authors. *Joint Conference on Digital Libraries (JCDL)* 2007, Vancouver, BC, Canada.

[15] Latour, B. 2007. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

[16] Latour, B, Jensen, P, Venturini, T, Grauwin, S, and Boullier, D. 2012. The Whole is always smaller than its parts. In *British Journal of Sociology*.

[17] Venturini, T. 2010. Building on faults: how to represent controversies with digital methods. *SAGE Journals*. December 5, 2010.

[18] VIVEIROS DE CASTRO, E. 2002. *A Inconstância da Alma Selvagem e Outros Ensaios de Antropologia*. São Paulo: Cosac & Naify. 552 pp.

[19] Berry, D. 2011. The Computational Turn: Thinking About the Digital Humanities. In *The Digital Humanities: Beyond Computing*. v 12.

[20] Blondel, V.; Guillaume, J.; Lambiotte, R.; Lefebvre, E. 2008. Fast unfolding of communities in large networks. In *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10), P10008 (12pp) doi: 10.1088/1742-5468/2008/10/P10008. ArXiv: http://arxiv.org/abs/0803.0476

[21] Grauwin, S. 2011. *Exploring Social Phenomena with Complex Systems Tools: The Journey of a Physicist in an Interdisciplinary Playground*. ENS Lyon, thesis.

[22] Viveiros de Castro, E. 2010. Intensive Filiation and Demonic Alliance. In *Deleuzian Intersections: Science, Technology, Anthropology*. Oxford: Berghahn.

[23] Gomes, L.; Almeida, V.; Almeida M. J.; Castro, F.; Bettencourt; L. 2009. *Quantifying Social and Opportunistic Behavior in Email Networks*. Advances in Complex Systems 12(1): 99-112.

[24] Barabási, A.-L.; Albert, R. 1999. *Emergence of scaling in random networks*. Science 286, 509–512

[25] Simondon, G. 2010. On the mode of existence of technical objects: Third part. (N. Mellamphy, D. Mellamphy, & N. B. Mellamphy, Trans.). London. Retrieved November 15, 2012, from http://www.academia.edu/539031/Gilbert_Simondon_The_Essence_of_ Technicity

[26] Viveiros de Castro, E. 2013. *La Mirada Del Jaguar: Introducion al Perspectivismo Amerindio*. Tinta Limon. Buenos Aires.

[27] Viveiros de Castro, E. 2012. "Immanence and Fear: Stranger events and subjects in Amazonia". In *HAU: Journal of Ethnographic Theory*. Vol 2 (1): 27-43.

[28] Vincenzo, N. 2008. *Modularity for community detection: history,perspectives and open issues.* Found at: http://supernet.isenberg.umass.edu/fulbright-catania/workshop-talks/nicosia-nagurney-daniele-workshop.pdf. Last access: 06/06/2014.

[29] Wagner, R. 2012. Facts force you to believe in them; perspectives encourage you to believe out of them. An introduction to Viveiros de Castro's magisterial essay. In *HAU: Journal of Ethnographic Theory*. Vol (1): 11-4.

Visualization of Gaze Tracking Data for UX Testing on the Web

Róbert Móro
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16
Bratislava, Slovakia
robert.moro@stuba.sk

Jakub Daráž
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16
Bratislava, Slovakia
xdarazj@stuba.sk

Mária Bieliková
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16
Bratislava, Slovakia
maria.bielikova@stuba.sk

ABSTRACT

Visualizations on the Web can help users to understand complex concepts, such as when too many objects of possible interest are present. For the purpose of evaluation of their usability, gaze tracking data represent a valuable source of information. These data are themselves complex, time-varying and in large quantities, thus posing challenges on their manipulation and visualization. We propose an infrastructure for collection and visualization of the gaze tracking data from dynamic Web applications. Its main purpose is to support researchers in UX (user experience) testing of their proposed interfaces (and visualizations). In the paper, we provide a user study on the usability of the infrastructure and compare it to existing solutions.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*evaluation/methodology, user-centered design*

General Terms

Design, Experimentation, Human Factors

Keywords

gaze tracking, infrastructure, visualization, UX testing, areas of interest, web

1. INTRODUCTION

For a picture (a visualization) to be worth a thousand words, it has to have a clear message that is easily understandable by the users (receivers of the message). However, visualizations nowadays are usually not only static pictures, but require often complex interaction with the interface elements, such as filtering values, selecting ranges (e.g. time, price, etc.), zooming or navigation. In addition, this interaction is in many cases carried out in the Web environment with dynamically generated, or even streamed content. Evaluating proposed visualization, its usability and the overall user experience (UX) can be, therefore, an uneasy task.

There are many questions that can be of importance during evaluation, such as: How much time do the users spend looking at the visualization and how much time interacting with the interface? In what order do they receive the information? Do they read the accompanying text? Does the pattern change when we change a particular element (its position, design, etc.)? In order to answer these questions, it is not enough to rely on the indirect or implicit forms of feedback, such as position of a mouse cursor, clicks or scrolling. We need to evaluate what the users are actually looking at.

For this purpose, we can utilize gaze tracking technology that is becoming more affordable for the researchers and the ordinary users alike. Existing solutions have, however, often only limited support for the Web-based dynamic applications. In this paper, we propose an infrastructure for gaze tracking data collection and visualization focusing on the Web environment. We developed a prototype that can transparently work with gaze tracking devices from various manufacturers and supports multiple browsers. We provide an empirical evaluation of the proposed infrastructure and its visualization capabilities for UX testing and compare it to some of the existing solutions.

2. RELATED WORK

Eye tracking has been applied in many user studies in the recent years. With lowering price and increasing availability of low-end models, it is becoming possible to have eye-trackers not only in UX laboratories, but also in end-users' notebooks. It opens up new possibilities for types of interactions and adaptation, i.e. personalization of the applications to the users as noted in [1]. The authors verified that adapting the displayed ads on a website based on gaze data resulted in significant increase of users' attention.

Adaptation of visualization based on gaze data was proposed in [5]. The authors compared two types of visualization, namely bar chart and radar graph on fourteen tasks of differing type and complexity. In addition, the participants' personal traits (cognitive abilities), such as perceptual speed or visual working memory have been tested. They were able to correctly classify the task's type, complexity and the users' cognitive ability based on the gaze data and selected areas of interest, thus showing, that there are distinct differences in patterns and interaction styles worth of adapting to the users.

Individual differences in gaze patterns and behaviours were

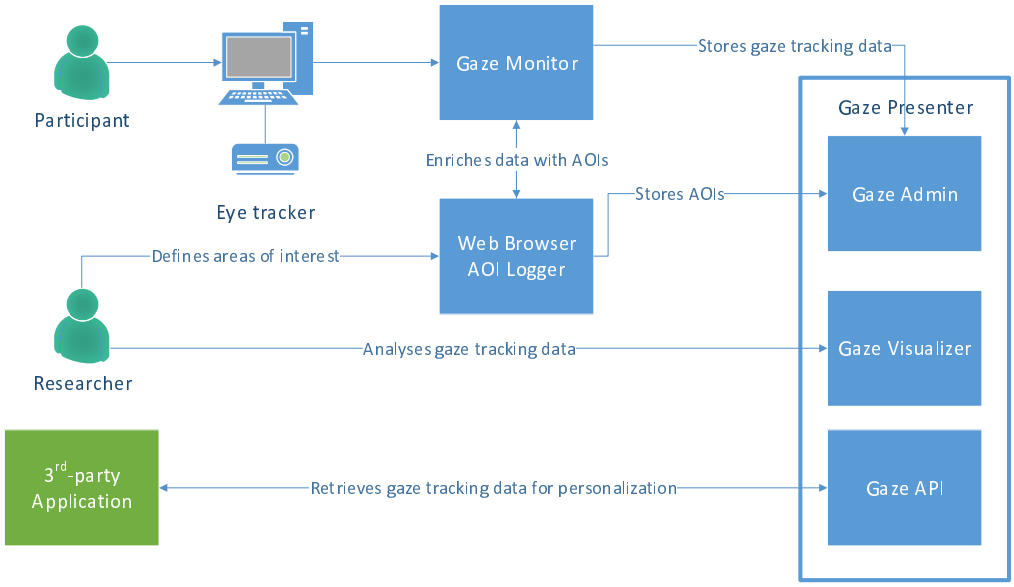


Figure 1: Conceptual design of the proposed architecture.

observed in [2] as well. Eye tracking has also been utilized in a user study on visualization of faceted interface [3]. The authors were interested in finding out, whether the users do not use facets just because they are shown to them. Therefore, they automatically hid (collapsed) them. Using the eye-tracker they verified that the faceted interface was used heavily in both cases (when visible as well as when hidden) with no significant difference in gaze patterns.

In order to be able to effectively evaluate areas of interest, we need to be able to track them throughout dynamically changing content. An algorithm for this purpose was proposed in [4] focusing on the tracking in video content.

However, tracking areas of interests on the Web usually requires different approach as the content can change completely, although it is still the same element (area of interest). According to our knowledge, it is still largely unsupported by the existing eye-tracking software.

The Eye Tribe[1] that promises a cheap tracker comes with no software, only with API for developers. On the other hand, *Tobii Technologies*[2] offers a *Tobii Studio* that comes with a full support for planning user studies, tracking, visualization and evaluation. However, it works only with Internet Explorer and areas of interest can be added only as static rectangles or polygons which is unusable with dynamically changing Web content. The best support for Web 2.0 seems to have *Nyan 2.0*[3] solution by *Eye Gaze, LC Technologies*. It can recognize different overlays and also visualize Web navigation paths. Areas of interest are, however, still defined as polygons. In addition, most of the existing solutions try to roll-out the Web pages to account for scrolling. This is, however, not enough for many modern applications, which can have different elements with their own scrollbars (e.g. Facebook with its chat, activity stream etc.).

Other problem with existing solutions is support for only one tracking device, i.e. multiple users cannot be tracked at the same time with exception of *Eyeworks* software by *Eyetracking*[4] when combined with their *Quad* server solution. Even so, the existing solutions for gaze data collection and visualization are developed by the eye-trackers' manufacturers and therefore, they are closed to one particular eye tracker brand and cannot be extended to work with devices from other manufacturers.

3. INFRASTRUCTURE FOR GAZE TRACKING

In order to address the problems of existing solutions discussed in the previous section, we propose an infrastructure for gaze tracking focusing on the dynamic Web applications. Its conceptual design can be seen in Figure 1. It consists of three main components, namely *Gaze Monitor*, *Web Browser AOI Logger* and *Gaze Presenter*, which in turn comprises of *Gaze Admin*, *Gaze Visualizer* and *Gaze API*.

Researchers define the areas of interest (AOI) using the *Web Browser AOI Logger* which are then stored on the server. They can set-up the whole experiment using the *Gaze Admin*, which is a part of *Gaze Presenter* component.

Then, the participants can connect using the *Gaze Monitor*, which communicates in the background with the eye-tracker, collects the gaze tracking data and sends them to *Web Browser AOI Logger* for enrichment. The data are enriched with the XPath[5] of the element the user (i.e. participant) is looking at, based on the coordinates supplied by the eye-tracker. The URL of the current website is added as well. Enriched data are sent by the *Gaze Monitor* at

[1] https://theeyetribe.com/
[2] http://www.tobii.com/
[3] http://www.eyegaze.com/eyegaze-analysis-software/
[4] http://www.eyetracking.com/
[5] http://www.w3.org/TR/xpath/

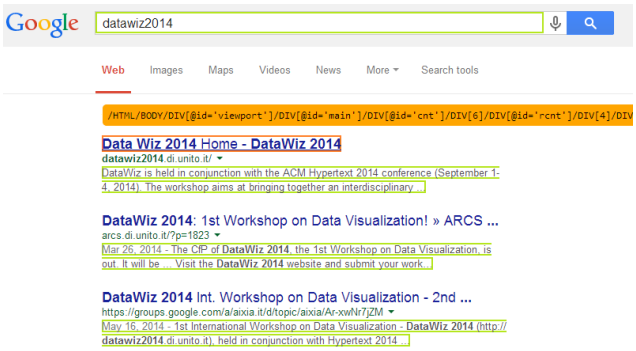


Figure 2: HTML elements highlighted during definition of areas of interest. Green ones have already been added (note that every snippet is a part of an area of interest definition), orange is highlighted upon mouse hover and can be selected by a mouse click.

specified time intervals to the *Gaze Presenter* for persistent storage.

They can be viewed and analysed by the researchers using the *Gaze Visualizer* component. The data can be also retrieved using the provided *Gaze API* and then manipulated by the third-party applications.

The individual components are further described in the following sections.

3.1 Gaze Monitor

Gaze Monitor connects to an eye-tracking device to receive gaze data from it. In order to transparently support devices from various manufacturers, we have implemented our own library that serves as a façade to the actual eye-tracker's API. Currently, we support devices from two manufacturers, namely *Tobii Technologies* and *The Eye Tribe*. In addition, we provide our own gaze data simulator that enables developers and researchers to develop applications for the eye-tracker without having one; gaze is simulated by the position and movement of the mouse cursor. Because it uses our provided library, the applications developed and tested with the help of the simulator can consume the simulated gaze tracking data as if they were from the actual eye-tracking device (i.e. using the same API calls).

The Gaze Monitor stores gaze data from the tracker in a queue. It communicates with the our provided browser extension - Web Browser AOI Logger, sends it the queued data and receives the enriched data. These are sent to the server in specified time intervals.

3.2 Web Browser AOI Logger

Web Browser AOI (Area of Interest) Logger is realized as an extension to the web browser. Its main functionality is to enrich data from the Gaze Monitor. Currently, we support both Google Chrome as well as Mozilla Firefox browser. The gaze tracker data contain normalized coordinates which are recalculated in order to identify the specific HTML element of the displayed Web page. The element is identified by its unique XPath.

Figure 3: XPath string. It can be customized by deselection of the specific path's elements (in gray).

The extension is also used to define areas of interest on the Web page, which is in more detail described in section 4.1.

3.3 Gaze Presenter

The data sent from the Gaze Monitor are collected by the provided server application, i.e. the Gaze Presenter. It enables data collection from multiple connected users at once. We use two databases for storing the data; SQL database for storing the information about experiments (projects, sessions, users, areas of interest) and NoSQL document-based database *RavenDB* for storing the enriched gaze tracking data in JSON format. One of the considerations when choosing the data storage was velocity of the incoming data; the eye-tracker' frequency is (based on the actual model) at least 30Hz meaning that we have approximately 100,000 new data records per each hour's worth of tracking.

The collected data can be accessed and visualized by the users using the provided Web interface. In addition, we provide an API for third-party applications that can consume collected data (i.e. what users are looking at which elements at what time) and e.g. adapt (personalize) the visualized information based on the users' gaze, i.e. what they are (not) looking at. Thus, the gaze tracking can be used not only for the purpose of evaluating the interface (visualization), but can be considered as a form of implicit user feedback. This way, it can help to model interests of the users more precisely.

4. VISUALIZATION OF GAZE TRACKING DATA

Visualization of gaze tracking data is crucial for its understanding and usage for evaluation of the user interfaces. Complexity lies in the data's velocity, multidimensionality and time variability. We can significantly reduce the computational requirements, when we include only data for specific areas of the tested Web page that are of an interest for us (so-called areas of interest). Thus, instead of computing fixations for all the elements, we can do it for a handful defined by the user.

4.1 Definition of Areas of Interest

We enable users to define areas of interest (AOI) using our browser extension. After activation, the elements in the Web page are highlighted upon mouse hover (see Figure 2). After the highlighted element is clicked on, the pop-up appears, in which it is possible to customize the selected area (name it, described it) or to change the XPath (see Figure 3) to suite the user's specific needs.

It is, thus, possible to choose not only the actual clicked element, but e.g. every paragraph with the same parent, or

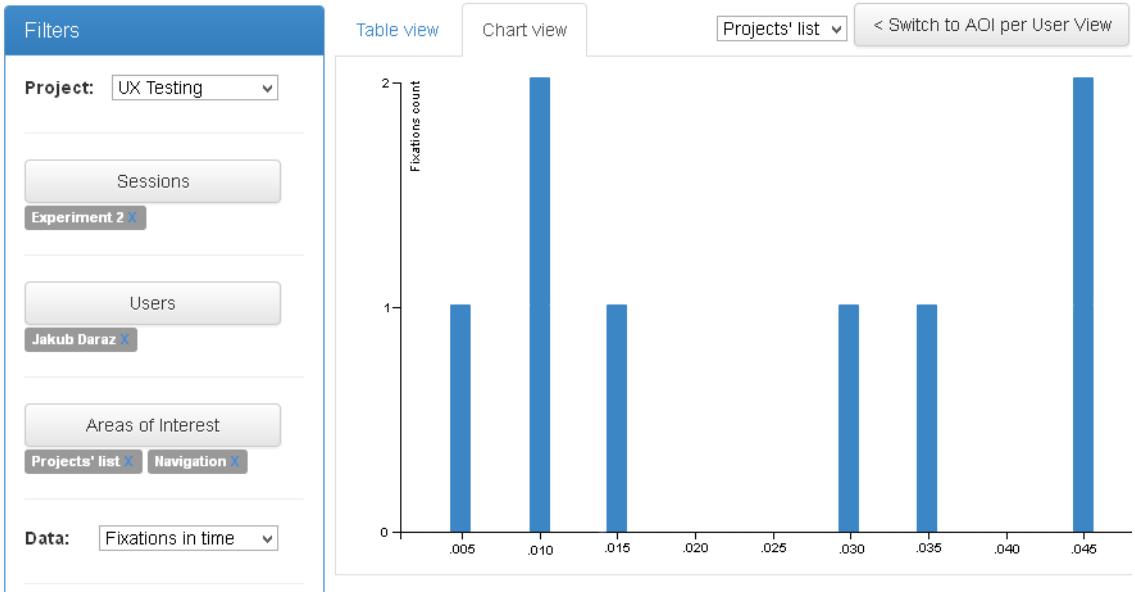


Figure 4: Visualization of fixations in time for the selected user and area of interest.

every element with the same class, etc. This can be used with advantage for the dynamically generated Web pages, where we do not know exact element's path, but we can identify it by its relative position within the HTML DOM structure or by its other attributes. It also enables users to include to an area of interest elements which are generated on the fly and are therefore not present at the time of area of interest definition, but share the same attribute value.

4.2 Visualization of Metrics

The eye-tracker tracks the position and movements of each eye separately; however, we are interested, what a user is looking at (which is rarely two things at once). Therefore, we calculate the gaze position as an average of the two eyes' coordinates. In addition, the tracker is not always precise and the gaze can seem to oscillate around a specific point, when the user actually looks at the same point the whole time. We use several smoothing techniques to account for this, especially a *moving average technique* by averaging N consecutive gaze coordinates from a moving window. The users (researchers) can also specify minimal time threshold for fixation, i.e. for how much time (e.g. 500 ms, 1 s, etc.) the user has to look at the area for it to count as a fixation. This way, we can filter out events, when the user moved gaze through the element without actually fixating on it.

The cleared data can be accessed and visualized by the users using the provided *Gaze Visualizer* component. Currently, we support the following metrics:

- *Number of fixations* - it counts, how many times the users looked at the specified areas of interest during the duration of the whole session
- *Dwell time* - similar to the first metric, but instead of the number of times the users' gaze entered the areas of interest, it aggregates the spent time (how long the users looked at the areas of interest during the whole session)
- *Fixation in time* - it shows, how the fixation count changed over the time of the experimental session (see Figure 4)

The users can aggregate and compare the data from multiple sessions, users and for multiple areas of interest using the provided filtering options. Data are shown in tabular view as well as visualized in the form of charts using the *D3.js*[6] library. The charts can be exported and saved to disk.

5. EXPERIENCE WITH THE PROPOSED GAZE TRACKING INFRASTRUCTURE AND ITS USABILITY

In order to evaluate our proposed infrastructure, we carried out a user study with four participants. We chose participants who had previous experience with eye tracking in Tobii Studio, so that they could compare the functionality of the both systems.

The participants' task was to set-up an experiment using our infrastructure, then collect the gaze data and lastly, to visualize and evaluate it. At the end, we asked them to fill in a questionnaire evaluating the different features.

The participants rated highly the provided functionality of defining the areas of interest. It was also rated as intuitive and easy to understand (4.25 on average from a five-point Lickert scale). However, we observed problems with editing the XPath string, namely the participants did not intuitively find out that it is customizable. After explanation of how it works, they appreciated the flexibility. One of the participant suggested that he would be interested to define not

[6] `http://d3js.org/`

only a single area of interest as a combination of different elements (e.g. each result on search engine's results page), but also to explore the differences in gaze patterns with individual elements within this area of interest group.

The participants found the experiment easy to set-up, although they had in some cases problems to understand the difference between a project and its sessions. As to the visualization, it was again rated very positively (4.25 on average), even though we currently provide only visualization of the three metrics. On the other hand, these metrics are ones of the most often used as we also verified in the reviewed literature (they were used practically in all of the related works reported in this paper). The participants missed the most possibility of creating heat maps and fixation sequences (how the gaze moves from element to element).

Compared to Tobii Studio, the participants appreciated the flexibility of defining the areas of interest, support of multiple browsers and multiple concurrent users as well as possibility to manually set the preferred minimal length (threshold) of fixations. On the other hand, they lacked audio and video recording and support of data inputs other than gaze, such as mouse clicks (left and right button), scroll events, etc. They would also appreciate the possibility to export the data or to clean it within our application.

Lastly, two participants would use our solution alone and two in combination with others, such as Tobii Studio, mainly for the lack of audiovisual recording. Overall, we find the feedback positive and encouraging for future development.

6. CONCLUSIONS

In the paper, we proposed an infrastructure for collection and visualization of gaze data focusing on the dynamic Web applications. Our main contributions are:

- visual definition and support of dynamic areas of interest, the content of which as well as size and position can change over time
- support of multiple browsers and eye-trackers from different manufacturers by providing a unified and easily extensible API
- collection and automatic evaluation of the gaze data from multiple concurrent devices and users

We realized a prototype of the infrastructure and carried out an user study in order to gain feedback to its functionality and usability. Based on the collected user feedback described in previous section, we plan to provide heat maps as well as fixation sequences visualization in the future. More importantly, we would like to enhance the data manipulation techniques, such as cleaning the data, selecting time ranges, zooming in and out, etc.

Currently, it is possible to automatically annotate the gaze data based on the fixations within the areas of interest defined by the users. However, the users may wish to add other annotations of different types either manually or automatically based on a set of predefined rules. It can be in a form of events, e.g. someone entered the room during the study, the participant looked away, the user study moderator provided a guidance, etc. These events represent useful metadata that can further explain the collected gaze data and provide new insights. In addition, it would be interesting to segment the data based on these events or compare the changes in gaze patterns or behaviour (e.g. before giving guidance and after it).

In order to support this kind of annotations, we have to solve several (also) visualization issues, namely visualization of gaze data stream in real-time and adding the annotations to a single point in data or a range. The easy to understand and intuitive visualization of the associated annotations in the data in the process of evaluation is also an open problem.

In addition, it is very likely that the eye-trackers will be a part of end-user devices in the near future. This will allow usage of gaze data as one of the implicit feedback factors of users' interest. When we combine our provided *Gaze API* with the events in the form of annotations, it can support new ways of personalized interactions on the Web.

7. ACKNOWLEDGMENTS

This work was partially supported by grants No. APVV 0208-10 and VG1/0971/11 and it was created with the support of the Research and Development Operational Programme for the project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the European Regional Development Fund.

We would like to thank our colleagues who participated on the development of the presented prototype, namely Dominika Červeňová, Lukáš Gregorovič, Michal Mészáros, Róbert Kocian, Martin Janík and Kristína Mišíková. We thank also the *Tobii Technology* for kindly providing us with the eye tracker as well as Tobii Studio for evaluation purposes.

8. REFERENCES

[1] F. Alt, A. S. Shirazi, A. Schmidt, and J. Mennenöh. Increasing the user's attention on the web. In *Proc. of the 7th Nordic Conf. on Human-Computer Interaction Making Sense Through Design - NordiCHI '12*, pp. 544–553, NY, USA, 2012. ACM Press.

[2] S. T. Dumais, G. Buscher, and E. Cutrell. Individual differences in gaze patterns for web search. In *Proc.of the 3rd Symposium on Information Interaction in Context - IIiX '10*, pp. 185–194, NY, USA, 2010. ACM Press.

[3] M. Kemman, M. Kleppe, and J. Maarseveen. Eye tracking the use of a collapsible facets panel in a search interface. In *Proc. of the 17th Int. Conf. on Theory and Practice of Digital Libraries - TPDL '13*, pp. 405–408, Berlin, Heidelberg, 2013. Springer.

[4] F. Papenmeier and M. Huff. DynAOI: a tool for matching eye-movement data with dynamic areas of interest in animations and movies. *Behavior Research Methods*, 42(1):179–87, Mar. 2010.

[5] B. Steichen, G. Carenini, and C. Conati. User-adaptive information visualization. In *Proc. of the 2013 Int. Conf. on Intelligent User Interfaces - IUI '13*, pp. 317–328, NY, USA, 2013. ACM Press.

TweetViz: Following Twitter hashtags to support storytelling

Lorena Lucas Regattieri
University of Alberta
Edmonton, AB
55 27 99767590
regattie@ualberta.ca

Ryan Chartier
University of Alberta
Edmonton, AB
recharti@ualberta.ca

Jennifer Windsor
University of Alberta
Edmonton, AB
jjwindsor@gmail.com

Geoffrey Rockwell
University of Alberta
Edmonton, AB
grockwel@ualberta.ca

ABSTRACT

How can visualizations of massive amounts of information be made more useful for data journalists? The availability of large amounts of publicly available user generated content is opening new opportunities to study social, cultural, and communications phenomenon. Computer assisted analysis now makes it possible to explore the relationship between nodes and text without having to choose between data size and depth. To create a visualization technique that would allowed us to reveal the network of actors and the main themes hidden in a large dataset, we had to work in a method of inquiry for social sciences. Based on the actor-network theory (ANT) we explored a dataset extracted from Twitter in order to map relationships and indicate new possibilities for journalists by discovering main themes around a hashtag, this way we interpret a layer of text multiple times, analyzing the nodes in its many attributes. Beyond the boundaries of 140 characters, this approach can succeed as it reproduces and reveals the dynamic connections contained in a collective phenomenon. In the last section, we demonstrate a prototype visualization that reveals behaviors and discourses within the large sample datasets. . We use the D3 visualization library to overlap related links and nodes to produce a comprehensible interactive visualization. Our model is interactive and allows us to identify part and whole pattern relationships constant with the three principles of information visualization: overview first, zoom and filter, then details on demand. This paper analyses networks from the perspective of ANT in order to create a visualization ready to support users when telling a story with data.

Categories and Subject Descriptors

D.3.3 **[Programming Languages]**: Language Constructs and Features – *abstract data types, polymorphism, control structures.*

D.2 **[Software Engineering]**: Design Tools and Techniques - *Flow charts, Object-oriented design methods, User interfaces.*

General Terms

Algorithms, Documentation, Performance, Reliability, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Design.

Keywords

Data journalism, Actor-Network Theory, design, social network analysis.

1. INTRODUCTION

A fair number of events and social phenomenon find themselves connected; they are caused by a range of parts of a complex puzzle interacting to each other. As a society [1], we came to recognize that nothing is isolated anymore. If not yet to consideration, the "global village" is even more a reality in the current state of living, where everything is linked. New understandings about society and community life are guided by a concept of "glocal" - something that translates the current sensation of being both, global and local [2]. The use of twitter data to interpreted human behavior is not news. Every day, more researchers are overcoming the issue of understanding social relations using text analysis and information visualizations tools. The availability of large amounts of publicly available user generated content is opening new opportunities to study social, cultural, and communications phenomenon. Computer assisted analysis now makes it possible to explore the relationship between nodes and text without having to choose between data size and depth [3]. As the volume of available information expands, it is becoming increasingly important for techniques to be developed that will allow for networks of information to be effectively summarized and navigated. The alternative—what has come to be known as the "hairball"—is becoming increasingly unwieldy and obfuscatory, no matter how many colour based filters are applied. To overcome the hairball we have developed a new visualization technique that allows us to reveal the network of actors and main themes hidden within traditional network visualizations of large datasets. In this paper we reveal this technique and our methods for producing it.

2. METHOD

The project began as a conversation about how to visualize large quantities of data and how this process could support data driven information. We made the decision to focus on hashtag and the Twitter conversations surrounding these hashtags. The conversation led a set of agreed upon features that are represented in the original sketch. The tool had to do two simple things: visualize the frequency of hashtags in a data set and allow the user

to click on specific hashtags and read the tweets associated with them. Every other feature we incorporated into the visualization serves one of these two purposes.

The tweets themselves were extracted using the Twarc[1] tweet scrapper. Twarc is a command line tool that takes a single search term (in this case the string 'rob ford'), queries the twitter API (Application Programming Interface), and the downloads all of the metadata associated with whatever tweets it finds. However, Twarc alone produces a large amount of unnecessary data. For every 140-character tweet that Twarc downloads, approximately five thousand characters worth of metadata is received. All told, we collected about twenty gigabytes of twitter data. The next step was to filter this data, for that it was built another scrapper, also in python, that would search this data and return in csv format all of the information needed. In this case hashtags, but many other attributes such as: geolocation, mentions, and url are also available. This dataset returned approximately one gigabyte of data. In order to filter the data further, we used an R script to split the csv files, format character codes and time stamps, as well as filter out every tweet that does not contain a hashtag. This reduced the dataset to two hundred megabytes. We then uploaded the entire remaining dataset to a MySQL database through a PHP script. The final step was to query this database for visualization in a JavaScript library: D3. For reaching out a visualization dashboard that could provide interactive information, D3 proved to be extremely useful. All told we employed seven different programs across six different programming languages in order to pre-process the data.

3. DISCUSSION

This paper situates the debate and challenges posed by the large amount information available online. In this matter, we begin with a context of critical questions on Big Data. Mathematicians, philosophers, sociologists, and many scholars from different fields of study are claiming "for access to the massive quantities of information produced by and about people, things, and their interactions."[4] Big Data is a term use for a large combination of datasets together. Following Manovich[3] observations on the issue, which puts Big Data near a researcher using a simple desktop, "we want to combine human ability to understand and interpret - which computers can't completely match yet - and computers' ability to analyze massive data sets using algorithms we create."

Data driven journalism is a field that brings together the interdisciplinary studies involving the provocations in big data and information visualization. According to Paul Bradshaw, data can be both, used in the production and distribution of information in the digital era and a tool with which the story is told. In journalism, like any source, data can be treated with skepticism; and like any tool, it "should be use with conscious of how to shape and restrict the stories that are created with it." [5]. Just to have an idea, the graphics department at The New York Times, has a group of about 30 people responsible for the information graphics and multimedia presentations, such as: reporting and writing copy, processing datasets, web development, drawing schematics, designing print pieces, and developing and creating the interface of multimedia projects. When selecting subjects to research, data analysis, and reporting, people from many backgrounds are doing data driven journalism, the fact that now the abundance of data has increased exponentially is a major challenge for the ones working in the area of visual storytelling.

Social network sites like Facebook, Instagram, and Twitter became a central component of sociability in our contemporary society. User generated content is a way to measure qualitative data, from the metrics on the success of a product inside the market to tracing the news about a natural disaster, social media delivers a massive amount of information everyday. In studies of network analysis, Twitter has become a broad database for quantitative and qualitative scholarly analysis. With user generated content and the flow of information, the microblog is the virtual space for peoples perspectives online [6]. Twitter is a rich environment for data scientists looking to investigate the issues of Big Data, social relation, and data visualizations. While sharing financial results in February, 2014, Twitter announced that its number of users has passed 241 million monthly active users. From the 215 million monthly active users, there is around 100 million daily active users, generating 500 million tweets per day. For qualitative research, Twitter offers a great strategy to segment a topic of interest, which is the hashtag (#). A topic is indicated through the composition of a hashtag and a keyword. This is the average practice in the use of "tags" when categorizing web content, anyone familiar with bookmarking will rapidly understand the importance of labeling certain tweets. A hashtag gain importance when the text has a high rate of retweets, meaning that a message is republished many times. This specific word will then reach Twitter's trending topics and achieve a level of importance. This will end up creating from time to time, specific topics of conversation between users. In qualitative research and for the purpose of this research, we will track the hashtags in order to examine its parts in the course of a news event.

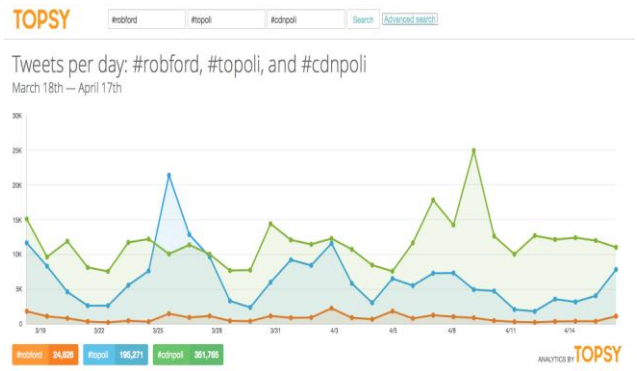


Figure 1. Frequency of hashtags overtime provides insights about topics: #robford #topoli #cdnpoli

Important and new questions emerge as we develop technical skills to overcome the "provocations" in Big Data, with computer assisted analysis it is possible to trace millions of opinions, ideas, feelings, and monitor those flux of information. Language, time, space, gain new features on the new method of information management. Thus, we need to think in new linguistic production associated with fast conversations on Twitter, for example, what would be the vocabulary during the course of an event, like a bomb explosion or a flooding? We can make these and many reflections analyzing the data extracted with the assistance of a

[1] Twarc was originally created to save tweets related to Aaron Swartz https://github.com/edsu/twarc

computer. In consequence, to tell stories based on these data visualizations.

The mapping controversies technique is a successful method to trace digital data. Cartography of controversies is a method created by Bruno Latour [7] and is broadly used in the communications field to map the debates around an specific object, subject, or event. This technique hinge on the idea that 'things' generate contested spaces, this way something new is produced following a large amount of material and subjective considerations. An Actor-Network-Theory (ANT) comprehension of events will move beyond the traditional dimensional image, between two or three common implications, extending to the meaning of the human factors, thus reducing the necessary to differ subject and object: "In a few words, when you look for controversies, search where collective life gets most complex: where the largest and most diverse assortment of actors is involved; where alliances and opposition transform recklessly; where nothing is simple as it seems; where everyone is shouting and quarrelling; where conflicts grow harshest. There, you will find the object of the cartography of controversies" [8].Considering Venturini's instructions to reach out for the controversy, we were lead to an investigation of an event that would be both, complex and big. A theme that would lead us to question the possibilities in the process of producing new visualizations, especially for data driven journalism. Knowing that we chose to pursue an empirical investigation within the course of news involving the Toronto mayor Rob Ford.

3.1 The story on the Rob Ford Controversy

A brief background about the case that explains the choice for data: starting in May 16, 2013, a series of reports about a video supposedly showing the Toronto mayor smoking from a glass pipe ends up circulating on the U.S media. Subsequently, media outlet Toronto Star also spread the news about a man their reporters claim in a video smoking crack. This is enough for the long controversy to begin. Since May, from denying allegations to new videos emerging from time to time in several news media, Rob Ford is an ongoing conversation on Twitter.

Building up from the theoretical references exposed above, we needed a dataset big enough to challenge us within the limits of back end and front end work with Big Data. With different themes underlying the discussion on Canadian and Toronto politics, the dataset extracted from Twitter around Rob Ford elaborates on how citizens are expressing their concerns on social, economics, and political issues in the society. The Rob Ford tweets set us up with long tail of conversations to follow, presenting us with a scenario demanding of critical thinking about information visualization. Moretti[9], Manovich[10], and Ruecker et al.[11] have drawn the attention of the literary research community to the value of visualization within the research process. Telling stories with data is about discussing theories of visual thinking and analytical design [12], however, it is also about engaging in a scholarly debate over the uses of a visual interface to investigate social data. We aim to bring together in our tool, an innovative method where anyone can quickly analyze, visualize and share information.

3.2 TweetViz: a tool to explore data[2]

In this section we demonstrate a prototype visualization that reveals behaviours and discourses within the large sample datasets. Our model is interactive and allows us to identify part and whole pattern relationships constant with the principles of Shneiderman's[13] visual information-seeking mantra: overview first, zoom and filter, then details on demand. We use the D3[3] visualization library to overlap related links and nodes to produce a comprehensible interactive visualization. In developing this technique we are untangling what would otherwise be "hairballs," aligning relevant information from the inside out, displaying clusters, outliers, patterns and trends, making visible to users "differences that make a difference". [14] An overview provides the gist of the data — the substance or salient aspects of the information and a perceptual shortcut. It is the 'macro' referred to when discussing micro/macro readings of information graphics: the texture of detail that we don't immediately need to direct our full attention to that cumulates into larger, coherent structures. Gist provides a summary of the data at a low cognitive cost for the viewer in terms of time and mental energy. The image (2) shows an early sketch of the concept, it was designed to allow comparisons to be made within an eye span and provides a general context for the entire dataset. The user then has a basis to draw on for further drill-down decisions.

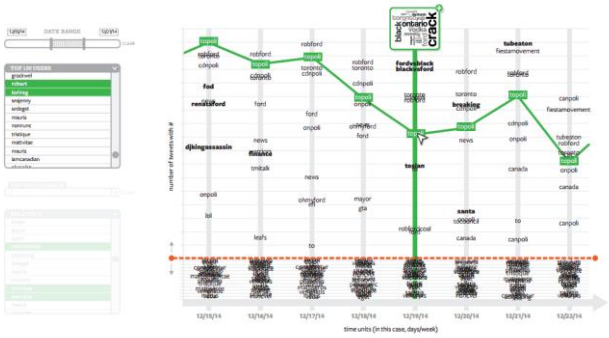


Figure 2. First sketch of the tool would display a word cloud for each day

Data visualizations excel at expressing comparative or relational aspects of data in order to highlight significant connections and identify patterns or trends. In the same way that mapmakers often focus on certain predetermined features of a landscape rather than depict an exact replica of an area from above, our first task in creating an overview of more than a million tweets was to consider which features were most likely to reveal relevant structures within, and context for, the data. When choosing a temporal framework for the visualization, patterns and trends (as evidenced by changes in the dataset such as new hashtag appearances, spikes in frequency and emergent word occurrence patterns) were revealed. It became possible to compare and contextualize data changes with real-world events. We chose hashtag frequency for the y-axis reasoning that it offered the broadest indication of tweet topic, and other means of drilling-

[2] TweetViz Prototype is available at http://analytics.artsrn.ualberta.ca/viz/hashtag.html

[3] D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS http://d3js.org/

down such as username and keyword search would then provide the viewer greater detail after. Highlighted hashtag occurrence over time, in the context of how often it appears, provides a macro view of a conversation arc over a given period. We also chose to highlight outliers — hashtags that only appear once in the data set — reasoning that they might provide a unique perspective from outside of occurring trends and patterns. After the broad strokes of the overview, the user can explore the data more closely. The 'zoom' Schneiderman referred to typically means changes in the scale of magnification — in TweetViz, it is semantic in nature. The user can move from a macro reading of the data to closer examinations of the text. In the original sketches, this is accomplished by either a small word cloud generated for a given hashtag each day, or in the tweets themselves in a second panel. Filtering is achieved with a date-range selector and a username and keyword search.

Figure 3. User can explore tweets by user or hashtag

A significant design concern for large data sets is dealing with occlusion: ensuring that the design inhibits visual elements overlapping as much as possible. In the early sketches, we designed a division between the 10 most commonly occurring hashtags and the rest of the hashtags in order to minimize overlap: when the slider bar is raised, the user can see all but the top 10 occurring hashtags in their relative (and often occluded) arrangement; when the bar is lowered, greater vertical space lessens overlap for the top 10.

The current visualization offers a toggle between relative and absolute views of the top 10 hashtags, and uses jitter — the slight, irregular movement of overlapping hashtags — to reveal overlapped elements at minute intervals. In the next paragraphs, we engage in the process of untangling the "Hairball" by building our own tool. The visualization dashboard consists of two screens. The first is a visualization of the relative frequency of each hashtag in the data. The larger the percentage of tweets that that hashtag gets used in the higher it appears on in the chart. Secondly, we also wanted to visualize the contents of these tweets. This is done in two ways. Firstly, the original design had a word cloud associated with each node, this word cloud is designed to offer an 'at a glance' insight into the content of the tweets represented by a hashtag. Secondly the user can click on a node and transfer the tweets in that node to the second screen. The viewer is simply a widget that allows the user to sort, filter, and read individual tweets.

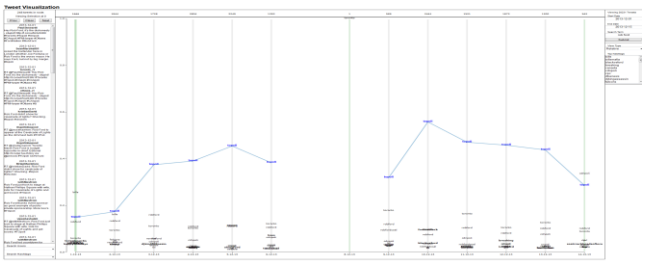


Figure 4. Visualization of the Relative Frequency of the hashtag #Topoli overtime

Due to the incredible quantity of tweets that twitter processes on a daily basis, the unique identification numbers assigned to each tweet was massive. Unfortunately, not every program handles large numbers in the same way, and due to the large assortment of programs in use, not all of this data was translated between languages perfectly. Another problem encountered is due to character encoding. Because twitter is an international platform, it is extremely lenient in which characters it allows. Unfortunately, due to the large amount programs and data formats used, not all of which allow by default the entire unicode character set, certain characters needed to be removed from the set (notably all newlines, carriage returns, and some foreign symbols I could not identify) and certain characters were lost in translation. An example of where this problem appears is in the 't' hashtag in the rob ford data set. Unfortunately, 't' is only a small part of the hashtag itself, but the rest does not render properly. Beyond the prototype stage a better solution to this project needs to be addressed. Request size also proved to be a problem. Javascript is a client side service, and in order for it to visualize properly the entire data set needs to be processed and transferred to the user computer. Unfortunately, due to the size of the project, these requests tended to overwhelm the earlier versions of the project. Early versions of the twitter viewer actually fetched the full text of every tweet it was analyzing. This was necessary because it was the easiest way to generate the word clouds dynamically. However, this soon proved to be too much for JavaScript to handle. Instead, we needed to preprocess all of the data on the server. Unfortunately, this meant that the word clouds needed to be generated outside of D3. Due to the difficulties to visualize, it was decided to cut the world clouds and only visualize the content through the tweet reader.

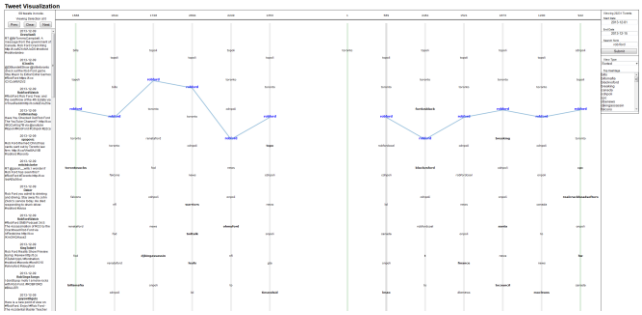


Figure 4. Visualization of the sorted Frequency of the hashtag #RobFord overtime

3.3 Reporting on issues and findings

Due to the incredible quantity of tweets that twitter processes on a daily basis, the unique identification numbers assigned to each tweet was massive. Unfortunately, not every program handles

large numbers in the same way, and due to the large assortment of programs in use, not all of this data was translated between languages perfectly. Another problem encountered is due to character encoding. Because twitter is an international platform, it is extremely lenient in which characters it allows. Unfortunately, due to the large amount programs and data formats used, not all of which allow by default the entire unicode character set, certain characters needed to be removed from the set (notably all newlines, carriage returns, and some foreign symbols I could not identify) and certain characters were lost in translation. An example of where this problem appears is in the 't' hashtag in the rob ford data set. Unfortunately, 't' is only a small part of the hashtag itself, but the rest does not render properly. Beyond the prototype stage a better solution to this project needs to be addressed. Request size also proved to be a problem. JavaScript is a client side service, and in order for it to visualize properly the entire data set needs to be processed and transferred to the user computer. Unfortunately, due to the size of the project, these requests tended to overwhelm the earlier versions of the project. Early versions of the twitter viewer actually fetched the full text of every tweet it was analyzing. This was necessary because it was the easiest way to generate the word clouds dynamically. However, this soon proved to be too much for JavaScript to handle. Instead, we needed to preprocess all of the data on the server. Unfortunately, this meant that the word clouds needed to be generated outside of D3. Due to the difficulties to visualize, it was decided to cut the world clouds and only visualize the content through the tweet reader.

In terms of visualization, crowding turned out to be the biggest problem in the visualization itself. Once the initial prototype was built on a small subset of the data, it became immediately apparent that some of the assumptions made in the original design were not true. The first assumption was that spacing between the top few hashtags would be relatively even. We could visualize the top hashtags as a relative percentage and use a slider bar to 'squish' all of the lower hashtags down allowing us to push them out of the way and focus on the higher percentage hashtags. In the Rob Ford data set, this is false, and in fact, the opposite is true. The top hashtags are completely dominant, and only the top three or so are actually visible on a relative scale with everything else squishing into the bottom. Instead, of using a slider bar to push the lower less important hashtags out of the way, it became apparent that we needed a way to focus in on the lesser hashtags and push the dominant ones out of the way.

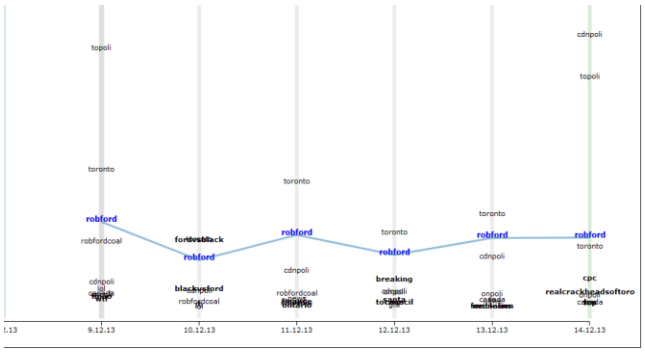


Figure 5. Issues when hashtags overlap each other

4. CONCLUSIONS

In short, our research builds up from a solid theoretical reference to visualize relationships in a network, the alliance between computing methods and the humanities consider quali-quantitative techniques that means more than overlapping statistical resources and ethnographic approach. We believe that information is only visible when the user can have the opportunity to click on, explore, discover, and share new findings. Data analysis can serve as technique to reveal the different structures of the same story and to provide new lens to see levels of information. When journalists use data to do their jobs they shift from being the first one to communicate to being the ones telling people what a certain progress of an event may actually mean. This tool can be appropriate by for journalists trying to visualize news and events, using data to transform something abstract into something everyone can understand and relate to the real events. With the curiosity to continue to think critically on how to display digital information and to explore data, for the future work we hope to overcome the issues with data encoding and crowding in our tool.

5. ACKNOWLEDGMENTS

Funding for the project generously supplied by Just What do They Do (JWDTD), Implementing New Knowledge Environments (INKE), and Social Science and Humanities Research Council of Canada (SSHRC).

6. REFERENCES

[1] Castells, M. 1996. *The Rise of the Network Society: The Information Age: Economy, Society and Culture, Volume 1*. Blackwell Publishers, Inc, Malden, MA.

[2] Wellman, B. 1999. *Networks in the Global Village: Life in Contemporary Communities*. Westview Press, Ed. Boulder, CO.

[3] Manovich, L. 2012. Trending: The Promises and the Challenges of Big Social Data. In *Debates in the Digital Humanities*. Minnesota, MI: The University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/15

[4] Boyd, D. and Crawford, Kate. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.In *Information, Communication, & Society* 15:5, 662-679.

[5] Gray, J. Chambers, L. and Bounegru, L. *The Data Journalism Handbook How Journalists Can Use Data to Improve the News*. O'Reilly Media.

[6] Wu, S. Hofman, J. Mason, W. and Watts, D. 2011. Who says what to whom on Twitter. In *International Conference On World Wide Web*, WWW'11,New York., NY.

[7] Latour, B. 2005. *Reassembling the Social: an Introduction to Actor Network Theory*. Oxford University Press. Oxford.

[8] Venturini, T. 2010. Diving in magma: how to explore controversies with actor-network theory. In *Public Understanding of Science* 19 (2009): 1-16.

[9] Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, London.

[10] Manovich, L. and Douglas J. 2009. Cultural Analytics. In *Plenary address at the Digital Humanities 2009 conference*. University of Maryland. June 22-25.

[11] Ruecker, S. Radzikowska,M. and Sinclair S. 2011. *Visual Interface Design for Digital Cultural Heritage: A Guide to*

Rich-Prospect Browsing. Farnham, Surrey: Ashgate Publishing,

[12] Tufte, E. 2001. *The Visual Display of Quantitative Information*, 2nd. Ed. Graphics Press LLC, Cheshire, Conneticut.

[13] Shneiderman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (Washington. IEEE Computer Society Press, pages 336-343) IEEE'96.

[14] Tufte, E. 2006. *Beautiful Evidence. Graphics Press*. Chesire, Connecticut:CT.

MarcoCivil: Visualizing the Civil Rights Framework for the Internet in Brazil

Lorena Lucas Regattieri
LABIC-UFES
regattie@ualberta.ca

Fabio Malini
LABIC-UFES
fabiomalini@gmail.com

Fabio Goveia
LABIC-UFES
fabiogv@gmail.com

Gabriel Herkenhoff
Labic-UFES
gabriel.herkenhoff@gmail.com

ABSTRACT

In this paper, we map the controversy surrounding the Marco Civil da Internet (Civil Framework for the Internet) in Brazil. Drawing on a Twitter dataset spanning from August 2012 to December 2013, this study uses a series of methods of data mining, processing, and information visualization to produce a historiography of collective actions related to the Marco Civil. The MarcoCivil platform at the "Digital Culture" website creaed initiatives to spread the discussions online: a Twitter profile @MarcoCivil (run by the administrators of the platform) and the MarcoCivil hashtag. To conduct the Marco Civil cartography we chose to work with the messages indexed to the MarcoCivil hashtag circulating on Twitter. In 2012 and 2013 Twitter became the online space in which cyber activists were most vocal. From October 2012 to January 2013 , we collected about 21.997 tweets related to Marco Civil, it was then that we noticed the presence of a controversy and a diversity of points of view in dispute. News reports in Brazilian newspapers during the discussion, little took into consideration the issues engendered in the struggle for approval of the law. By demonstrating with graphs the dispute between the different actors involved in this battle, we seek to contribute to the history of the approval of the Marco Civil. From telecommunications companies to politicians, our report show how history was made in the field of the internet human rights.

Categories and Subject Descriptors

D.3.3 [**Programming Languages**]: Language Constructs and Features – *abstract data types, polymorphism, control structures.*

K.4.1 **[Computer and Society]:** Public Policy Issues – *Ethics, Intellectual property rights, privacy, regulation.*

G.2.2 **[Numerical Anlaysis]:** Graph Theory - *Graph algorithms. Network problems.*

General Terms

Algorithms, Management, Measurement, Documentation, Performance, Design, Reliability, Experimentation, Human Factors, Languages, Theory, Legal Aspects, Verification.

Keywords

Civil Rights Framework for the Internet, journalism, data mining, social network analysis, complex networks.

1. INTRODUCTION

After an intense debate in 2007-08, the Office of Legislative Affairs of the Ministry of Justice, in partnership with the School of Law in Rio de Janeiro at the Getulio Vargas Foundation (FGV), initiated the collaborative construction of the proposal for a civil law framework for the Internet in October of 2009. The goal was to create legislation that defined "the legal responsibility for providers and users for the content posted on the Internet…[and identified] measures to preserve and regulate the fundamental rights of Internet users, such as freedom of expression and privacy".[1]

The Civil Rights Framework for the Internet in Brazil opposes the tendency to establish restrictions, convictions or bans on the use of the internet. The framework intended to determine clearly the rights and responsibilities regarding the use of digital media. The focus, therefore, is the establishment of a legislation ensuring rights, not a rule restricting freedoms. Between November 2009 and June 2010, the Marco Civil was developed through a uniquely open public process that allowed all Brazilian Internet users an opportunity to comment on its text. In the spirit of the bill's substance, civil society was empowered to collaborate with policymakers in order to make the bill reflective of public interest and prioties.

An initial draft drawn by legislators was posted Cultural Digital, an open platform where the public could submit and review suggested changes to the bill. Throughout an open debate, Marco Civil received over two thousand comments from academics, civil society organizations, technical experts, and private individuals. In 2011, the Marco Civil was submitted to Congress as Executive Bill 2126 and was given priority on the legislative agenda. Since then, the bill has become the subject of numerous controversies in the House of Representatives due to inflammatory issues such as network neutrality, privacy, freedom of expression, and copyright. The Bill has made it onto the agenda of the House of

[1] An English version of the bill is available at FGV http://direitorio.fgv.br/sites/direitorio.fgv.br/files/Marco%20Civil%20-%20English%20Version%20sept2011.pdf

Representatives eight times, but each time the vote has been postponed due to the lack of agreement among Members about crucial points in the Marco Civil.

Challenges in reaching an agreement have created an obstacle to the consolidation of a national-level regulatory framework for the Internet. Among other things, this immobility reveals a tension between the interests of businesses and the demands of civil society. Over the course of the bill's legislative history, the telecommunications lobby and content industries have been the driving force behind significant changes to the text. During this period, we have also witnessed a somewhat "schizophrenic" dynamic take hold of policymaking efforts concerning the Internet. While the Ministry of Justice created an innovative collaborative platform so that civil society could participate in the production of "The Bill of Rights for the Internet," it also saw broad mobilization around a bill that sought to combat all forms of crime on the Internet, especially financial crimes. Meanwhile, the Parliament endeavored to focus on criminal laws as a foundational aspect of Internet regulation in the country.

This strange situation persists today, as the copyright and telecommunications industries oppose free "peer to peer" exchange and net neutrality. This can be explained, in part, by the interests of public security forces, which after public protests in June 2013 (strongly articulated by the civil society through social networks) advocated establishing a longer required period for the retention of private communications data that could support the investigation of crimes and "deviations". The situation was compounded in the wake of the Edward Snowden leaks revealing the National Security Agency (NSA) spying other countries through PRISM. This struck a chord for Brazilian President Dilma Roussef, who subsequent to the leaks, proposed an amendment to the Marco Civil that would force foreign companies to host data on national servers. The proposal has proved highly controversial, due both, to the geopolitical implications it would carry and the technical complications it could introduce.

Within the approaval of the Marco Civil, the world turns the eyes to Brazil when it comes to Internet civil rights. The world celebrated the bill at the NETmundial – Global Multistakeholder Meeting on the Future of Internet Governance and at Arena Participative. At the Arena, we had the presence of important people discussing internet and human rights, such as Roy Singham (ThoughWorks), Julian Assange (Wikileaks) from the Ecuador Embassy, and Frank La Rue (ONU). The event that brought together representatives of governments and civil society in search of a letter of international principles for the Internet was considered the beginning of the process to discuss the internet policies in a global context. History was made, but it is crucial to understand the path to the approval of the Marco Civil in order to comprehend the struggles involved in the fight for Internet human rights.

2. METHOD AND GOALS

Latour[1] and Venturini's[2] mapping controversies technique is successful method to trace digital data. It is broadly used in the communications field to map the debates around a specific object/event. This is the theoretical foundation guiding our research; we used the cartography method to support us in the digging experience in the Twitter data. As an empirical template, Twitter served us for the purpose of:

- Map the network of controversies on the #MarcoCivil;
- Perform a semantic analysis of the expressions, hashtags, and controversial issues that circulated on Twitter under the #MarcoCivil hashtag.

We centered our analyses around two distinct periods:

- July - December 2012: The Marco Civil bill enters the agenda of the House of Representatives
- July - December 2013: Discussions about the bill resume at the House of Representatives.

In our network visualization, we chose to plot the network of retweets (RTs) that included the #MarcoCivil hashtag. Since RTs must be replicated by many individuals, RTs on Twitter indicate that a subject (represented by a hashtag) carried significant social relevance. We extracted data directly from the Twitter API, which allowed us to capture and store about 20,000 tweets produced by almost 10,000 profiles monitored in 2012.

For each tweet, we were able to log the tweet text, date, origin and destination of the tweet. The subsequent step after mining and processing is the data is the visualization of data. Using the open source tool Gephi[2], we sliced the data using different metrics, creating new graph visualizations for each metric. To support our semantic analysis of the data, we analyzed 5137 tweets to identify the political position of each actor in the debate on #MarcoCivil; the way Twitter profiles were expressing themselves in the network; the intention of the message; the themes it touched upon; and the controversy.

The second procedure was to analyze all the tweets, 21,000 in 2012 and 110,000 in 2013. For this, we used a data-mining tool called NAR_T[3], a python script developed within the Laboratory of Studies on Image and Cyberculture (LABIC). The script provides the following outputs:

- Most repeated words and hashtags.
- Most replicated tweets.
- Word clouds and hashtags.
- Co-occurring hashtags network.
- Most mentioned users.
- Number of tweets per user.
- Number of active users per day

After generating groups with Gephi, we extracted the profile names that built up each cluster in the network. When we processed the script with the "cluster_usernames" of each of the groups, we obtained the same outputs, but now we could analyze them by targeted group. This allowed us to investigate the unique positions surrounding the controversy of each of the groups identified.

[2] Gephi is an open-source software for visualizing and analyzing large networks graphs. Available at: *http://**gephi**.org*

[3] This script was created to parse tweets. It is available at https://github.com/ufeslabic/parse-tweets

3. DISCUSSIONS

3.1 General Observations of Marco Civil Network Dynamics

In August 2012, when the Marco Civil entered the voting agenda at the House of Representatives, the politics of this power struggle overflowed into the virtual universe, particularly on social networks. This chart represents the high level of participation on Twitter, especially, the days in which the bill was expected to be voted at the House of Representatives.

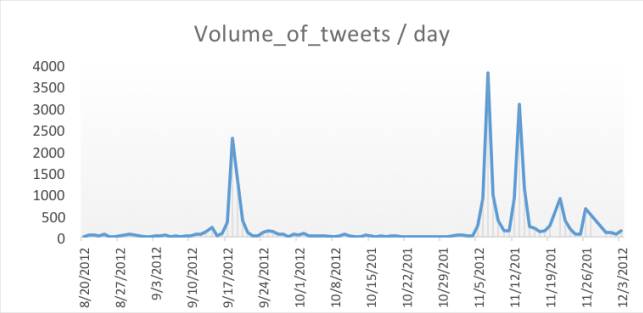


Figure 1. Number of Tweets per day with the hashtag #MarcoCivil on Twitter, from 21 August to 3 December 2012.

With the vote imminent, activists, parliamentarians, lawyers, specialists, businessmen, intellectuals, artists, government ministers and even President Dilma Rousseff used social networks to produce a broad debate on the subject. The buzz over the Marco Civil quickly became one of the longest standing controversies in the recent history of Brazilian politics. The increasing rate of publication of tweets directly correlates with increased political debate around the subject. The closer the House of Representatives was to voting on the legislation, the more activity we saw on Twitter under the #MarcoCivil hashtag. The representatives found themselves facing pressure from a broad range of channels: social networks, emails, blogs, and online media. Some party websites even underwent DDoS attacks. Digital expression around the issue became a strategy for activists. In many ways, these tactics exposed many politicians to public judgment, affecting their image among voters. This strategy has proven to be a key measure to the movements connected to the field of free culture and the most progressive deputies.

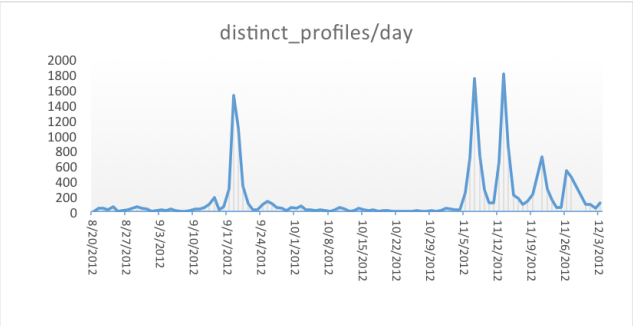


Figure 2. Number of unique users per day participating in the publication of tweets with the hashtag #MarcoCivil.

From August to December 2012, heightened publicity around the bill generated the mobilization of 16,072 different profiles, 22651 tweets and 5640 retweets (Figure 2). A variety of profiles and the volume of tweets eventually formed an interactive network with different common points of view on distinct aspects of the law (Figure 2).

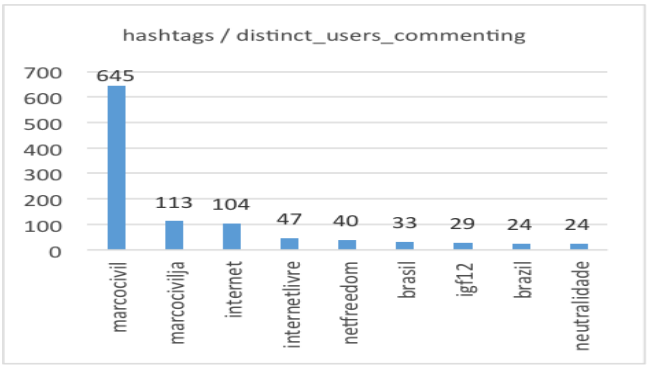


Figure 3. High rate Hashtag use with the hashtag #MarcoCivil.

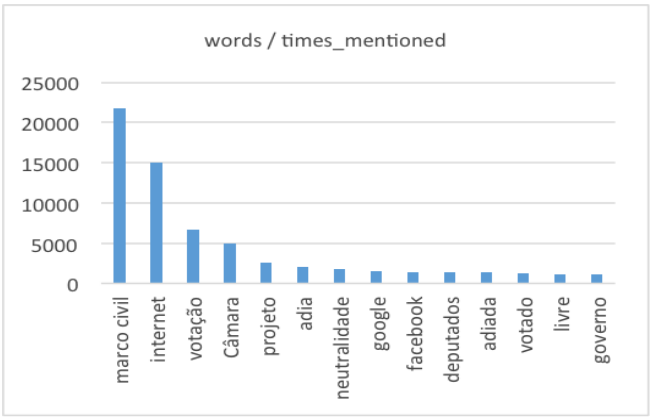


Figure 4. Word frequency within tweets mentioning the hashtag #MarcoCivil

Frequent use of the terms "vote" (votação) (6652), "postpones" (adia) (2065), "House"(câmara) (4941) and "bill of law" (projeto de lei) (2616)" suggested high levels of expectation that the bill would pass and a commitment, at least among a minority of users, to monitoring the long and tiresome journey of Marco Civil in the Congress. The anxiety around the bill was highlighted by the intense correlation of the hashtag #marcocivil with the #MarcoCivilJá (#MarcoCivilNow). The word “neutrality” and the #neutrality hashtag can be seen often in the dataset (Figure 3 and 4) suggesting it was the most commonly discussed subject in interactions between members of Congress and users tweeting about #MarcoCivil.

3.2 Marco Civil in 2013: the network is polarized and the privacy debate gains attention

In 2012, the difficult process in voting the bill 2126/2011, plus the numerous delays and changes in the course of the project, turned the social networks - notably Twitter - into a major platform for discussion about the Marco Civil. Activists, experts and concerned individuals began to debate the issue, seeking to defend their perspectives and understand the significance of the bill for the future of the Internet in the country. But with the failure to reach an agreement and the start of the municipal elections of 2012, the vote on the Marco Civil fell into oblivion,

eventually being suspended. In June of 2013, two critical events affected the trajectory of the bill: Public uprisings throughout the country and the first of the Snowden leaks. Protests over transit fare hikes, economic inequality and other "bread and butter" issues peaked in June, with some protesters referencing the bill and making it part of their messaging, both on and offline. At the same time, some activists began to argue against the creation of the civil framework for the Internet, claiming that the Marco Civil was a ploy made by the government to restrict Internet freedom. This questioning came up in light of the numerous arrests of Facebook page administrators from groups opposed to the government, in particular, Anonymous and Black Blocs [3]. Back then, videos from Anonymous began circulating claiming that Marco Civil was going to have the opposite effects: for them the intentions of Marco Civil were to control online content. Thus, a trend of polarization emerged while some continued to promote the bill, despite changes in the text that weakened user protections in the face of copyright restrictions, others began voicing opposition to the bill, arguing that it would lead to greater Internet censorship. The perspective of media outlets came out exactly between these two groups, as news feeds reflected the arguments of both sides. The emergence of groups that made radical critiques of the Marco Civil represented a fundamental shift in the debate on the subject. This change can be better understood when we undertake a semantic analysis of the network formed by these groups during this period of time.

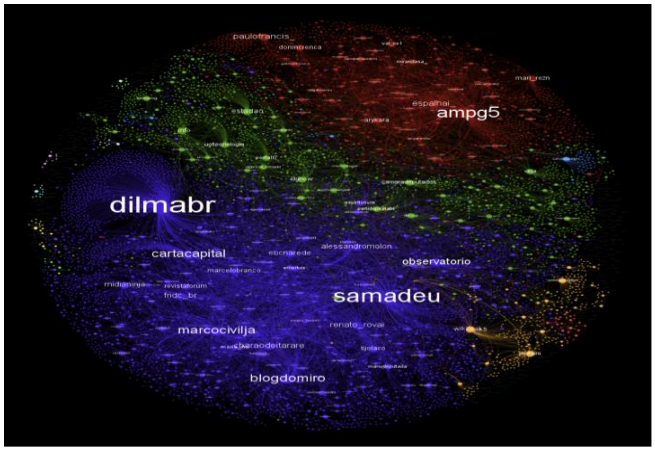


Figure 5. Network of profiles that participated in the debate on the Marco Civil from July to December 2013. In the Spotlight, profiles whose messages were most popular in the network.

The graph in Figure (5) shows the relationship established through retweets from profiles that between July 17 and December 31 that used the keyword "Marco Civil". To produce this visualization we processed the data with the high gravity scale to bring closer together those actors who had more connections with the group to which they belong. After this first step, we generated a statistic of modularity in order to visually emphasize each perspective by assigning each a different color. We used the metric of authority to give prominence to nodes that had both stronger and larger quantities of connections in the network, with the goal of finding those individuals who had a higher indegree in the Marco Civil controversy. All told, the final goal was to display those who received the highest number of RTs of other important actors in the network. For these groups, sharing messages creates links between the actors in the network and illustrates a force of attraction between them (a dynamic referred to as "gravity".) As an individual, typically (thought not always) shares ideas with those, which he agrees, individuals with similar opinions share content with and from each other, creating groups, which we call perspectives. There are four perspectives within the Marco Civil network:

• The purple network: individuals in favor of voting on the law (46.55% of the total network)

•The red network: individuals contrary to voting the bill (17.39%)

• The green network: media outlets and profiles specialized in law and civil rights (20.56%)

• The yellow network: foreign organizations that generally supported the proposal for a regulatory framework (4.1% of total)

4. CONCLUSIONS

Our study suggests that the free digital culture activists are the ones responsible for articulating the Marco Civil debates. Thus, social networks i.e. Twitter, prove to be a rich environment for the open debate. This network has become a major strategy to pressure the Brazilian Congress. In our study, we employed computer-assisted analysis through mining methods and data visualization in order to investigate our hypothesis. The outputs have proven that our hypothesis is correct, as our research displays several indications pointing to the centrality of the actions and pro-Marco Civil campaign coordinated by activists from Brazil and around the world. The days before voting on the Marco Civil by the House of Representatives were periods when Twitter profiles became highly mobilized in order to debate and press the Parliament on the approval (or not) of the Marco Civil. This demonstrates that the community formed around the hashtags remained attentive to the decision-making movement of Congress. On the other hand, it demonstrates how politics is creating a routine towards the emotional tone of networks, influenced by the chaotic flow of public opinion on the Internet.

5. ACKNOWLEDGMENTS

Funding for the project 'Mapping Controversies on the Internet: a scientific cooperation between researchers who analyze the relationship between Aesthetics, Power and Internet' generously supplied by National Council for Scientific and Technological Development (CNPq), National Academic Cooperation Program (Procad), Coordination of Improvement of Higher Education Personnel (Capes), Foundation of the Ministry of Education (MEC). Our thanks to the team at the Laboratory of Studies in Image and Cyberculture (LABIC) for the continues support.

6. REFERENCES

[1] Latour, B. 2005. *Reassembling the Social: an Introduction to Actor Network Theory*. Oxford University Press. Oxford.

[2] Venturini, T. 2010. Diving in magma: how to explore controversies with actor-network theory. In *Public Understanding of Science* 19 (2009): 1-16.

[3] Passos, N. 2014. *O Black Bloc e o papel das mídias sociais nas manifestações brasileiras de 7 de setembro de 2013.* Unpublished.

Explorations in Media Visualization

Everardo Reyes-García
University of Paris 13
99, av. Jean-Baptiste Clément
93430 Villetaneuse, France
everardo.reyes-garcia@univ-paris13.fr

ABSTRACT
In this contribution we explore an emergent approach to data visualization called 'media visualization'. The main characteristic of this practice is to take into account the content of visual media directly as a constituent part of the data visualization project. Media visualization employs and develops image processing techniques. It contributes to current efforts on the design of data visualization such as diagrammatical representations, spatial distribution of elements, combination of colors, or animated behaviors. In this paper we describe 'media visualization': principles, requirements and related work. We also show some examples of media visualization developed by us within the framework of visual analytics and media art.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *prototyping, screen design, interaction styles*.

General Terms
Design, Experimentation, Human Factors.

Keywords
Media visualization, visual representation, visual analytics.

1. INTRODUCTION
Research and development in data visualization (here understood as un umbrella term for associated notions such as information design, visual representation, or even hypermedia models) have gained popularity and acceptance for depicting discreet data in graphical form. Today, we see how some graphical models that once were restricted to particular domains become common and distributed. Models such as network visualizations (force-directed graphs, among others), treemaps, and streamgraphs are more and more present in diverse professional domains (newspapers, mass media, etc.)

Within this diversified context, the kind of data that is visualized deals most of the time to social records, transactions, preferences, hours, locations, connections, etc. There is also a considerable amount of valuable tools and resources to produce data visualization, ranging from scripting libraries (d3.js, sigma.js, etc.) and software applications (Tableau, Gephi, etc). However, the same cannot be said when we try to analyze and organize a corpus of complex data which includes visual media such as photographs, films, or any other digital images, broadly speaking. Of course, images are often used in visualization projects. The best examples are infographics and geographical maps, whose main role is often to contextualize and decorate statistical data. Yet it is not common to find a project that analyses and represents visual features of images themselves.

In this contribution we focus on media visualization as an emergent approach to take into account visual media as constituent part of a visualization project. After describing its primary goals and techniques, we will present some examples developed by us in order to reflect on our own experiences and to identify future work.

HT'14, September 1-4, 2014, Santiago, Chile.

2. MEDIA VISUALIZATION
'Media visualization' is an idea originated in 2005 and currently developed by the Software Studies Initiative [12]. It refers to the practice of analyzing visual media through visual media. In other words, it consists of making visualizations including the images being analyzed. In contrast to common data visualizations, where data is most of the time depicted as symbols and organized in diagrams, media visualization takes advantage of visual analytics and image processing techniques to construct visual spaces of the information analyzed.

In general, a project on media visualization involves two domains: digital image processing and information design. The first domain is useful to extract and measure visual features from a collection of images, while the second domain concentrates on the visual representation of the collection of images. A project on media visualization assists research on cultural analysis through the identification of patterns by means of visual analytics [3].

Images must be understood technically and plastically, and not exclusively from the figurative standpoint. Digital images are series of pixels with chromatic values arranged in a bi-dimensional matrix. The visible content of the image could be regarded from two perspectives: figurative or non-figurative (also known as plastic). Figuratively, the accent is on recognizing characters, objects, places, etc. On the contrary, the plastic strand considers images as fundamentally chromatic values, forms, and shapes. These three properties constitute the visual features of the image.

For us, visual features define the realm of materiality and objectivity of images. These features can be seized and quantified. In computer science, visual features are the operational units inside image processing procedures such as: analysis, extraction, classification, retrieval, visualization, and representation [8].

In the case of colors, the process of extracting and measuring visual properties implies storing in the database different values that represent chromatic information. We can use for example the HSB color model as the basis for measuring images. In one

column we can have its median hue value, in the second column its median saturation, and in the third column its median brightness. Of course, just as we decided to calculate the median value, we could also calculate the average, the mean, the standard deviation and other statistical measures.

In the case of forms and shapes, the associated data considers the visible area, particles, fragments, contours, distribution, and dimensions, among others [8]. Other measures could concentrate on features such as block differences and variations; entropy; Sobel edge detection; Adaptive Color Quantization; statistics on RGB channels, etc.

Besides visual features, each image can also be described semantically with metadata. The database can be enriched with categories: year, designer, photographer, creator, software used, place, technique, etc.

Once the database has been assembled, we are now in position to look for modes of representation of images. As we said, the idea is to make evident patterns to approach cultural analysis. Current media visualization techniques require digital images as input data in order to output a new, different, and processed digital image. So far, few techniques have started to delimit the practice of media visualization.

2.1 Image Pixelation

Image pixelation consists basically on obtaining the colors of an image and to represent them according to a discreet sequence of mask shapes. The mask shape is often a square (but could also be another geometrical figure such as circles or triangles) and its color is sampled from the original image and organized along its relative position to the image. The size of a unitary shape determines the degree of pixelation. A bigger size of shape implies the summarization of more colors from the visual area where it gets its values.

2.2 Image Averaging

Image averaging consists on stacking a series of images on top of each other at the same spatial coordinates. It implies that all images are present in the same visual space, but in order to observe visual patterns it is necessary to perform a statistical measure of visual features, otherwise only the last image of the series would be visible. A single procedure for image averaging would be to reduce the opacity of each image by n-times its percentage. Another technique would be to output an image where each pixel depicts the calculated measure in all the series of images.

2.3 Image Mosaic

Image mosaic, also known as image montaging, consists on ordering the corpus of images one after another in a sequential manner. Such as texts and grids, images are arranged in lines and columns. The ordering rule could be obtained from measures of visual features (for instance going from the brightest to the darkest), from metadata (for instance by year) or by order of appearance in the sequence (from the first to the last frame). The resulting image montage shows a rhythm of variations and transformations. In many cases it seems visual patterns are clearer when there is no space between columns and lines (i.e. images are only divided by their own size) and when all the images of the corpus have the same dimensions.

2.4 Image Slicing

Image slicing also presents the corpus of images one after another but there is a fundamental difference in comparison to an image mosaic. We call a 'slice' a thin part of an image, a region that slices it all along its X or Y axis. A slice does not show or summarizes the entire image, but only a delimited region. The size of the slice (how thin of thick it is) can be parameterized. For large collections of images, it seems thinner slices are the best option in order to depict variations and transformations of the entire corpus of analysis. The visual patterns then are observed by differences and variations in the regions generated.

2.5 Image Plotting

Image plotting is based on common types of 2D plots that use dots and lines to represent data along the X and Y axis. An image plot places, at the crossing coordinate of two values, the image corresponding to those values. So, for example, we can decide to plot images by 'year' on the X axis, while the Y axis would determined by the median brightness value. In this case, we can observe variations and evolution in time over the two scales.

2.6 Related Work

This brief review of emerging media visualization techniques emphasized two of its underlying domains: image processing and information design. Both domains have a history outside modern data visualization. For instance, image processing flourished in computer vision, computer graphics, and scientific visualization. Media visualization takes advantage of tools and techniques from these developments to create its own procedures. Currently, one of the main software environments to extract and measure visual features is ImageJ, which is open source and well-known among specialists of medical imaging [4]. Besides a series of scripts and software on top of ImageJ, other tools are QtImageProcessing, Mondrian (for statistical operations), scripts for MathLab, and VisualSense.

Regarding information design, we observe a close relationship between media visualization and contemporary art. In fact, some existing techniques can be approached from media art. Pixelation, for example, is related to 'pixel art', as introduced by Goldberg and Flegal in 1982 to describe the new kind of images being produced with Toolbox, a Smalltalk-80 drawing system designed for interactive image creation and editing [2]. Image averaging is related to the work of Sirovich and Kirby on 'Eigenfaces' in 1987 [11], and more recently, to Jason Salavon, who has produced a series of images by averaging 100 photos of special moments [10]. For image mosaics, Brendan Dawes presented 'Cinema Redux' in 2004, a project aimed at showing what he calls a visual fingerprint of an entire movie [1]. His main idea was to decompose an entire film into frames and then to arrange them as rows and columns. And image slicing can also be seen as a remediation of slit-scan photography. Among other prominent slit-scan photographers, William Larson produced, from 1967 to 1970, a series of experiments on photography called 'figures in motion'. The trick was to mount a thin slit in front of the camera lens to avoid the pass of light into the film. Thus the image is only a part of an ordinary 35mm photograph.

To conclude this section, we think 'media visualizations' have been focused so far on visual media: photographs, comics, magazine covers, album covers, film photograms, etc. But we know there are other types of media which are not visual, or not only visual. There is still work to do on audio, gestures, performance, tissue, garments, objects, furniture, industrial design, architecture, virtual worlds, and hybrid and multimodal media. Among other issues, there is more research to be done in analyzing and representing sound as sound (sonorisation rather than visualization) and objects as objects. In any case, it is important to remember that media in digital form implies the

transformation of another media form. An image of a painting or an album cover is a representation of the physical object; and an image of a digital image is its encoding, reproduction, compression, modification, and rendering.

3. EXPLORATIONS IN MEDIA VISUALIZATION

In this section we present our work on media visualization. The following projects have been developed mainly as research and experimental practice; like tools for reflection. While putting in practice existing techniques and methods for cultural analysis, we try to explore new forms of representation and interaction. One of our strategies has been the exploration of the aesthetics of digital information through visual disruptions, that is, by reconfiguring the expected functional mode of visual representations [6].

For the following examples we concentrate on information design and presentation formats. The first example is more related to the exploration of shapes, and the second to the exploration of colors. The presentation format is studied as a constraint of designing the resulting media visualization. We know early projects in media visualization were static, in the form of a single high-resolution image, which are useful for print and exhibitions. Likewise, first interactive explorations of image collections were done in large tiled computer displays (such as the 287-megapixel HIPerSpace at Calit2). But if the presentation format is web-based, we must face the challenge of smaller screens and the speed of network connection. Similarly, if the presentation format is a 3D shape, the challenge is on rendering and interacting with 3D models for the web or even on printing them for analog and manual analysis.

Presentation formats have their own conventions for explaining visualizations. For media visualizations, we often see texts, lines, arrows and other indicators that assist the identification and labeling of patterns. In a conference poster, for example, the designer can manually layout elements and design symbols and diagrams to improve comprehension. In a video narrative, titles and sound facilitate making sense of patterns. In a web-based context, recent projects combine different views and information processing techniques such as filtering, searching, and sorting.

3.1 Motion Structures

'Motion structures' is an ongoing project initiated in 2011 [7]. The idea is to convert an animated video sequence into a 3D digital model with the intention of revealing patterns of time and shape. The mode of interaction with a 3D object allows for different ways of digital exploration: orbiting around, zooming in and out, and immersive views inside the model.

Our process puts special attention on shapes above other visual features. To create a motion structure we use a script we wrote for ImageJ. Basically, the operations require decomposing an animated video sequence into a series of separated image files, which are then manipulated as an image stack. Then, several image processing techniques occur under the hood: converting images to 8-bit format, subtracting background, and rendering the stack as 3D shape.

With motion structures we intend to represent the spatial and temporal transformations of a moving image sequence. The obtained 3D shape encodes the changes of the objects in a frame: the different positions, the movement traces, and spatial and temporal relations. The way in which we can interact with an object is not limited to ImageJ. The model can be exported and later manipulated in other 3D software applications such as Maya, Sculptris, or MeshLab. Furthermore, it is also possible to export a motion structure for the web or to physically print it, however both techniques require destructive 3D model processing, i.e. reducing geometry by simplification, decimation or resampling. For technical details, a motion structure exported from ImageJ has an average of 500,000 vertices and more than 1 million faces, which is a very large amount compared to an optimized 2000-face model for the web, loaded with the library three.js.

The current constraints of the exploration of motion structures in web-based environments and as printed objects can be seen as a similar path to the evolution of the representation of movement. Pioneers such as Etienne Jules Marey and Eadweard Muybridge first represented movement with pictures and images themselves, but later Frank Gilbreth abstracted the traces of movement and created diagrams made out of lines. At that moment, artists got inspiration from both types of representation with the intention to explore a vocabulary of symbols, myths, and psychic processes.

So experimental projects on media visualization contribute to the design of data visualization in two manners. First, through the abstraction of shapes, traces and patterns, it permits discovering diagrammatical representations, spatial distributions of elements, and combinations of shapes. Second, through the inclusion of images and the design of exploratory and immersive experiences, it provides insights for investigating animated behaviors, combinations of colors, mix of media, and graphical indicators to improve the comprehension of patterns in non-figurative productions.

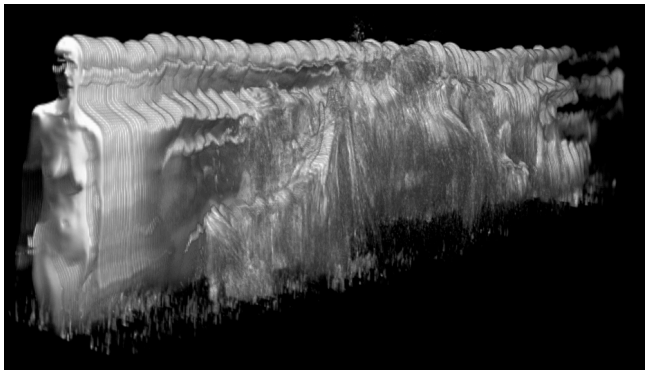


Figure 1. Motion structure from Bill Viola's Intimate Work

3.2 Web-based Media Visualizations

Our last example is an exploration in web-based media visualization. We developed 'RockViz' with the intention to produce web-based image mosaics and image plots [9].

For this project we gathered data about the most significant Rock albums according to AllMusic.com. Data was obtained Rovi, the data service behind AllMusic. The total amount of records was 1994, ranging from Blues to Alternative Rock and Heavy Metal. The metadata collected was about the artist/group name, album title, release date, and album cover image. Contrary to 'motion structures', for RockViz the principal visual feature to explore was the chromatic values.

The first step for our media visualization was to download all the images and make them available locally, so we could measure their chromatic features. We used ImageMeasure, a script for Image, to measure hue, saturation, and brightness values. Then, we used Open Refine to handle data, but more importantly to apply mathematical formulae to the measure of images and dynamically calculate their Cartesian position.

The image mosaic was ordered according to, first, median of hue; second, median saturation, and third, median brightness. To facilitate the exploration of the dataset, we added a filter engine that acts upon years, artist name, and album title. Finally, to make a little faster the loading of images, we produced two versions of each image: one is scaled to 100 x 100 px. and the other to 500 x 500 px. The small version is used for visual representations and the larger appears when the user clicks on an image, so she can observe more details of a single cover. Of course, a deeper study should consider larger dimensions of images but this was the largest resolution provided through AllMusic.

For image plots, we calculated spatial positions according to measures of visual features. We decided to use Open Refine to dynamically generate the HTML for each image because of two reasons. First, Open Refine supports algebraic and trigonometric operations so we could restrain the visual area to fit a resolution of 1024 x 768 px. Second, we originally used JQuery and the function getJSON to communicate with a JSON database, but the loading time is very slow for more than a few hundreds of images.

While an image plot requires translations of scale, for instance years into maximum width in pixels (in our case 1024 px), we also experimented with different representations inspired by geometric figures. We used the main formula for Cartesian-Polar transformations. Our first exploration, Figure 2, draws images around polar coordinates, taking values from median hue and median saturation, resulting into a chromatic circle-like visualization. Figure 3 disrupt this formula to investigate how images could be plotted according to different figures.

Web-based media visualizations contribute to information design in recalling the need to adapt large amounts of information to small screens. Moreover, it raises questions on making efficient time-consuming operations for transferring data files. But considering the web as presentation support also demands to reconsider the value of early developments by the hypermedia community and their potential implementation with contemporary web technologies. We are thinking specially in the xanalogical model, where visualizations of transclusions are depicted in a 3D environment [5].

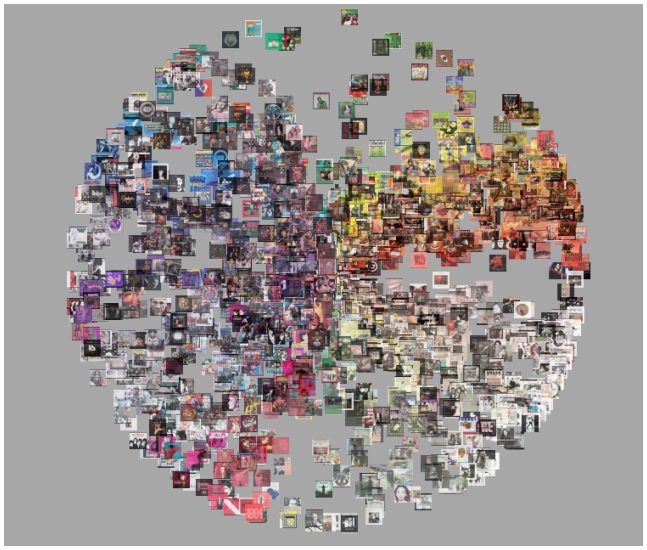


Figure 2. Experimental interactive web-based plot of images

Figure 3. Experimental interactive web-based plot of images

4. FUTURE WORK

So far, the primary goal behind our media visualizations has been to practice and experiment with existing techniques. The interpretation of data and the explanation of visual cultural patterns have been put aside momentarily, but this is precisely one of the clues for our future work. We expect to use our visualizations as teaching resource but also to collaborate closely with historians, filmmakers, media artists, musicologists and other domain experts.

On the other hand, we believe there is still much to do at the level of interactivity. Research on hypermedia functionalities needs to be done for web-based visualizations. In the same line, models of representation also require to be tested and experienced. While text-and-number-based visualizations meet an explosion of models, diagrams, demos, libraries, etc., some of them simply are not suited for visual media. We believe there are two main domains where we can get valuable insights: media art and scientific imagery. For the former, artists often challenge our tools and our common viewing experience; they practice could be regarded as very innovative. For the latter, we must remember that digital images are not exclusive to design, arts and art history, a wide range of different disciplines use them as well: geography, astronomy, medical imaging, mathematics, physics, chemistry, biology, etc.

5. CONCLUSION

In this paper we have made a brief review of the emergent approach of media visualization: its main principles and requirements. We identified this approach mainly at the coupling of image processing techniques and information design. Today we can list a short gallery of media visualization techniques and projects that start settling guides and practices. In order to enrich current research and development on media visualization we observed that two domains are particularly interesting: on the one hand, the heritage of hypermedia functionalities, systems, abstractions, and models for web-based projects. On the second hand, experimental media art and software from other disciplines

equally related to images (others than media studies and art history, for example sciences).

In the last part of our contribution we presented two explorations on media visualization. First, 'motion structures' an experiment on transforming an animated video sequence into a 3D digital model with the intention of revealing patterns of time and shape. Second, 'RockViz' a web-based media visualization comprising almost 2000 rock album cover images and its visualization through experimental image plots.

Further work should be conducted in the design visual indicators to improve the comprehension of media visualizations, which are often non-figurative and difficult to seize. At the same time, the non-figurative character of resulting processes can be seen as a move towards the symbolism of abstractions. By abstracting shapes, traces and patterns, new models emerge and can be applied to other domains.

6. REFERENCES

[1] Dawes, B. 2004. *Cinema Redux*. Online: http://brendandawes.com/projects/cinemaredux

[2] Goldberg, and Flegal, R. 1982. "Pixel Art" in *Communications of the ACM*. Vol. 25. No. 12. December. New York: ACM Press.

[3] Manovich, L. 2008. *Cultural analytics: analysis and visualization of large cultural data sets*. White Paper. Software Studies Initiative. Online: http://softwarestudies.com/cultural_analytics/cultural_analytics_2008.doc

[4] National Institutes of Health. 2014. *ImageJ: image processing and analysis in Java*. Online: http://imagej.nih.gov/ij/

[5] Project Xanadu. Online: http://www.xanadu.com/

[6] Reyes, Everardo. 2012. Disrupting 3D Models in *Proceedings of the 3rd. Computer Art Congress*. Paris: Europia.

[7] Reyes, E. 2013. *Motion structures*. Online: http://ereyes.net/ms/

[8] Reyes, E. 2013. "On Visual Features and Artistic Digital Images" in *Proceedings of the ACM conference VRIC'13*. Laval, France.

[9] Reyes, E. 2014. *RockViz: visualizing the 2k. most significant rock album covers*. Online: http://ereyes.net/rockViz/

[10] Salavon, J. 2004. *100 Special Moments*. Online: http://salavon.com/work/SpecialMoments/

[11] Sirovich, L. & M. Kirby. 1987. "Low-dimensional procedure for the characterization of human faces" in *Journal of the Optical Society of America*. Vol. 4. No. 3, pp. 519–524.

[12] Software Studies Initiative. 2014. Online: http://lab.softwarestudies.com/

Blind Browsing on Hand-Held Devices: Touching the Web... to Understand it Better

Waseem Safi[1] Fabrice Maurel[1] Jean-Marc Routoure[1, 2] Pierre Beust[1] Gaël Dias[1]

[1] University of Caen Basse-Normandie – UNICAEN
[2] National Superior Engineering School of Caen - ENSICAEN
14032 Caen- France - +33(0)231567336 +33 (0)231452722

{waseem.safi, fabrice.maurel, jean-marc.routoure, pierre.beust, gael.dias}@unicaen.fr

ABSTRACT

Navigating the Web is one of important missions in the field of computer accessibility. Many specialized techniques for VIP (Visually Impaired People) succeeded to extract the information displayed on digital screens and succeeded to transform this information in a linear way either into a written format on special Braille devices, or into a vocal output using text to speech synthesizers. However, although this success, screen readers failed to transform the 2-dimentional structure of the navigated web page; despite many researches confirm that perception the structure enhances web navigation and memorization. In this paper, we propose a new technique aimed to enhance the VIP ability to navigate the Web by affording a "first glance" web page overview. This technique focuses on improving non-visual vibro-tactile access to web pages on touch-screen devices, based on extraction and re-organization the structure of texts and graphical elements for web pages, reformatting and converting automatically these visual structures and textual information into vibrating pages using a graphical vibro-tactile language.

Categories and Subject Descriptors

[H.5.2] [**User Interfaces**] Graphical user interfaces (GUI), Haptic I/O, User-centered design, User interface management systems.

General Terms

Algorithms, Design, Human Factors, Languages.

Keywords

Visually impaired people, non-visual access, graphical vibro-tactile language, visual structures, textual information.

1. INTRODUCTION

In October 2013, the world health organization estimated that the number of VIP in the world is 285 million, 39 million of them are blind, and 246 million of them have low vision. The organization defined four levels of visual functions depending on the international classification of diseases, which are: normal vision, moderate visual impairment, severe visual impairment, and blindness[*].

VIP depend on screen readers in order to deal with computer operating systems and computational programs. One of most important and desired targets by VIP is navigating the Web, considering the increased importance and expansion of web-based computational programs. Screen readers present some solutions to navigate the Web, either by transforming a web page into a written Braille, or into a vocal output. Some screen readers installed on touch devices transform a web page into a vocal-tactile output. But there are some drawbacks for these proposed solutions: on the one hand, the Braille techniques are costly, and only few number of VIP have learned Braille (in France, there are about 77 000 visually impaired people and only 15 000 of them have learned Braille -from statistics published in September 2011-) [**]. On the other hand, transforming the information of a web page into a vocal format might not be suitable in public and noisy environments. Finally most of Braille solutions are not suitable for mobile devices [1]. In addition to these drawbacks, the most important one is the failure to transform the 2-D web page structure, because as reported by many authors, perception the 2D structure greatly improves navigation efficiency and memorizing the information because it allows high level reading strategies (rapid or cursory reading, finding or locating information,...) [2]. Our work focuses on developing and evaluating a sensory substitution system based on vibro-tactile solution which may solve the mentioned drawbacks; where we study how to increase the VIP perception of a 2-D web page structure, and how to enhance their techniques to navigate the Web on touch-screen devices. This suggested solution is very cheap comparing with prices of Braille devices, and also it could be more efficient in noisy and public environments comparing with vocal-tactile solutions. Our contribution is three-fold:

- Designing a Tactile Vision Sensory System (TVSS) represented by an electronic circuit and an android program in order to transform light contrasts of touch-screen devices into low-frequencies tactile vibrations,

- Running a series of experiments with blind persons in order to validate our hypotheses, and

- Analyzing many navigation models and tactics of blind persons.

The paper is organized as following: firstly, in section 2 we view the state of the art for VIP targeted technologies, and then we describe theoretical and methodological approaches for the new proposed technique, these approaches will be presented in section 3. The first pre-tests achieved with blind persons will then be described in section number 4, and analysis of results will be presented in the fifth section. Finally, in the last section the conclusion and perspectives will be proposed.

* http://www.who.int **http://www.opc.asso.fr/

2. STATE OF THE ART FOR VIP TECHNOLOGIES

Current products for VIP such as screen readers depend mainly on speech synthesis or Braille solutions, such as ChromeVox [3], Windows-Eyes [4], and Jaws (Job Access With Speech) [5]. Braille displays are complex and expensive electromechanical devices that connect to a computer and display Braille characters. Speech synthesis engines convert texts into artificial speech, where the text is analyzed and transformed into phonemes; these phonemes are then processed using signal processing techniques.

Some screen readers can support a tactile feedback when working on touch devices, such as Mobile Accessibility [6], Talkback [7] for Android, and VoiceOver [8] for IPad. Many of these products propose shortcuts for the blind user to display a menu of HTML elements existed in the web page, for example headers, links, and images. But, the main drawback of all these products is that they transfer the web page information into a linear way, and without any indication for the global web page structure (2D layouts).

Many researches tried to enhance the way by which VIP interact with web pages, such as [9], that proposed a tactile web navigator to enable blind people to access the Internet. This navigator extracts texts from web pages, and sends these texts to a microcontroller responsible of displaying the text in Braille language using an array of solenoids.

A tactile web browser for hypertext documents has been proposed by [14]. This browser renders texts and graphics for VIP on a tactile graphics display, and it supports a voice output to read textual paragraphs and to provide a vocal feedback. The authors implemented two exploration modes, one for bitmap graphics, and another one for Scalable Vector Graphics. Main drawback of this proposed system is that it needs a pin matrix device, which is expensive and cannot be integrated with handled devices.

Another interesting model called MAP-RDF ("Model of Architecture of web Pages") [10] proposed a method to improve the accessibility to visual information for blind persons. This model allows representing the structure of a web page, and provides the blind users with an overview of the web page layout and the document structure semantics. The main drawback of this model is that it could be applied only on well structured web pages which contain meta-data, so it could not be applied to most web pages which rarely contain meta-data. This model transforms the HTML elements to graphical symbols as illustrated in figure 1.

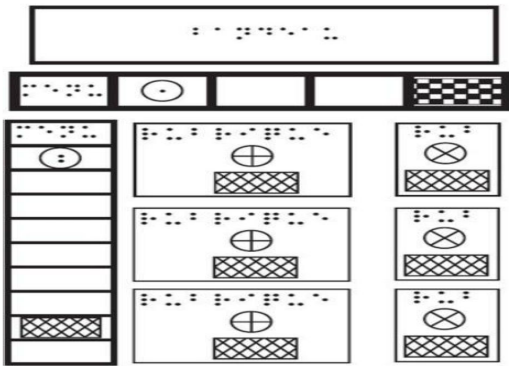


Figure 1. Representation of a web page by MAP-RDF model, figure extracted from [10].

In figure1, we notice many symbols; each one represents an HTML element. For example, the symbol represents a menu of items. And the symbols and represent texts with cold and hot colors.

Tactos is a perceptual interaction system suggested by [11], it consists of three elements: 1- tactile simulators (two Braille cells with 8 pins) represent a tactile feedback system, 2- a graphics tablet with a stylus (represents an input device), 3- computer [12], as shown in figure 2.

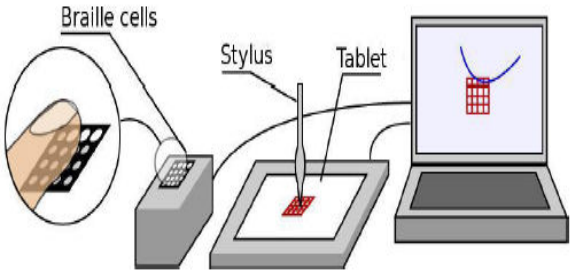


Figure 2. Different devices used in Tactos (Computer, graphics tablet, tactile simulators), figure extracted from [12].

The graphics tablet and the stylus allow the user to explore graphical contents on the screen such as circles, rectangles, and characters. While the user explores the contents, the system transforms pixels under the stylus into tactile stimulation on the Braille cells.

30 prototypes of Tactos have been released, to be used by a lot of users in many domains. Tactos has been successfully used to recognize simple and complex shapes. The device has been also used in geometry teaching domain in an institution for visually impaired and blind children. Tactos also allowed psychology researchers to propose and develop new paradigms for studying perceptions and mediated communication of blind persons [12].

3. PROPOSED TECHNIQUE

First glance could be defined as the ability -in a blink of an eye- to understand the document layout and its structural semantics [1]. We aim of our work to increase the ability of visually impaired persons to understand the web page 2-dimentional layout in order to enhance their tactics to navigate the Web. A commercial tablet connected to a vibro tactile set-up is used for that.

The first phase in our model is to extract visual structures in the navigated web page, and to convert these visual blocks to zones (segments) for facilitating the navigation in later phases. We achieve this phase depending on a hybrid segmentation method. Then the system will represent on the tablet screen the extracted visual elements as symbols using a graphical language (this language is under-development). The third phase is to browse these graphical symbols depending on size of the used touched-screen device, and then in the fourth phase, our system provides a vibro-tactile feedback when the blind user touches the tablet. The intensity and the frequency of the vibration depend mainly on gray level under the finger. A tablet (Asus Model TF101 with Android operating system) has being used for our tests.

In this paper, we focus only on the fourth phase which specializes in giving the user a vibro-tactile feedback by transforming light

contrasts of touch-screen devices into low-frequencies tactile vibrations.

To achieve the desired system, we have designed an electronic circuit which controls two micro-vibrators placed on the hands. A Bluetooth connection with an android tablet allows controlling the vibration intensity (Amplitude) of vibrators. An Android dedicated program on the tablet displays an image on the screen and detects where the user touches the tablet screen. The gray level of touched points is transmitted to the embedded device in order to control the vibration intensity. At the moment, only one micro-vibrator was used for pre-tests described in this paper.

Figure 3 illustrates the designed electronic circuit, and the used vibrator.

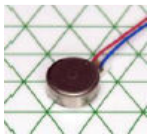


Figure 3. (a) The used micro-vibrator

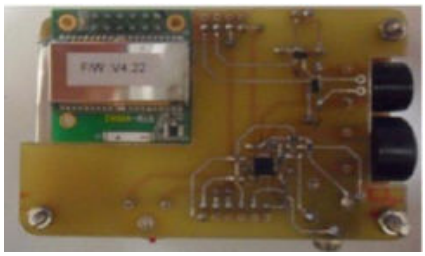


Figure 3. (b) The embedded system

Figure 3. (a) The used micro-vibrator: the range of vibration frequency goes from 20 Hz up to 260 Hz. (b). The embedded system designed in the lab with the Bluetooth module.

4. PRE-TESTS PROTOCOL

4.1 Objectives of Pre-tests

Our objective of the designed protocol is enhancing the ability of VIP to recognize the 2-D structure of a web page. In order to test the prototype mentioned in section 3, we designed some images contain different structures (detailed in section 4.2), and we tested the prototype firstly on 15 sighted persons (their eyes were closed) [13], and later on 5 blind persons. Testing the protocol on sighted and blind persons gave us a more understanding of tactics and strategies achieved by sighted and blind persons to navigate the designed structures. This will be useful in designing the desired graphical vibro-tactile language (all results are detailed in next sections).

4.2 Designed Protocol for Vibro-Tactile Access

Each experiment (either for sighted or blind persons) consists of 4 ordered phases of training (learning task), and four ordered phases of evaluation (evaluation task). All the experiments were filmed, and the designed program stocked many parameters in log files (coordinates X, Y, pressure on the screen, and the time at each touch). Figure 4 presents the 4 images of training phases, and figure 5 presents the 4 images of evaluation phases.

Figure 4 (a). Image a (Training task).

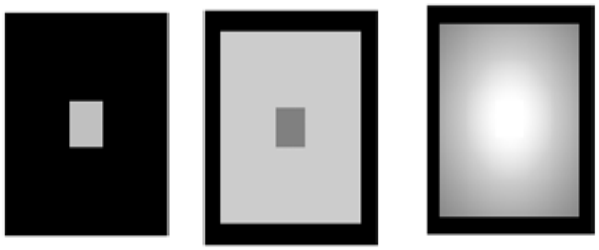


Figure 4 (b). Images b (NT2), c (NT3), d (NTG) (Training task)

Figure 4. Images of training task.

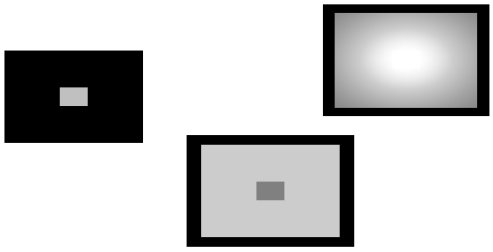


Figure 5 (a). Image a (Evaluation task).

Figure 5 (b). Images b (IDP1), c (IDP2), d (IDP3). (Evaluation Task).

Figure 5. Images of evaluation tasks.

In the training task, each user discovered firstly graphical elements in each image presented in figure 4 (images a, b (NT2), c (NT3), d (NTG)), and users were informed about names of graphical elements. The name of each image NT2, NT3, NTG, indicates how many transitions are necessary to access the square center, for example NT2 proposes 2 transitions to access the center of the square.

The evaluation task consists also of 4 phases, the first one allows to discover the image 5.a and to name each square inside it, then next phases are about images 5.b, 5.c, and 5.d, where we asked users to discover contents of each image, then to describe these contents, and to redraw discovered elements inside each image. We chose these images depending on following considerations:

- Image 5.a contains all squares on which users have trained in the training task, so it could test the ability to memorize and to distinguish the shapes.

- Image 5.b contains 3 rectangles with matched sizes and with vertical order, and the image 5.c contains 3 rectangles with different sizes and many relations of directions, so testing images 5.b, and 5.c could test the ability of distinguishing sizes, and distinguishing relations of directions.

- Image 5.d contains different shapes (a rectangle and a polygon), so it could test the ability to distinguish different shapes in the same image.

- The tested images contain examples of expected results of the segmentation process, so success of distinguishing these shapes by blind users could be an indicator of their ability to distinguish results of segmenting web pages.

The results of pretests with sighted persons were already published in [1]. Table 1 presents some results of the experiment for images NT2, NT3, NTG (the time required to distinguish graphical elements and the number of errors for the 15 sighted persons. The users have been asked to name the shapes in figure 5.a, and for each shape, we evaluated the number of correct and incorrect answers). In table 1, number or errors represents the number incorrect answers.

Table 1. Results of experiments of sighted persons for images NT2, NT3, NTG. (Table extracted from [1])

Shape Name	Average Time in Seconds	Number of errors
NT2	28	2
NT3	36	6
NTG	22	3

We notice from table 1 that the lowest number of errors is assigned to image NT2, and the largest time and max number of errors is assigned to image NT3.

5. RESULTS OF EXPERIMENTS WITH BLIND PERSONS

5.1. Experiment steps:

The test performed with each one of the 5 blind persons consisted of following: personal and technical questions, explanations of the test objective, a training task, and finally an evaluation task. The approximated average time for the test for each person is about 1 hour.

5.1.1 Personal and technical questions

Before starting the tests with the 5 blind persons, we asked them to support us with some information about their age and date of their blindness. Table 2 summarizes answers of personal questions.

Table 2. Personal information of the blind persons

User-ID	0	1	2	3	4
Age (Years)	63	67	59	56	36
Sex	Male	Female	Male	Female	Female
Date of the blindness	Since Birth	since 32 years	since 25 years	Since 10 years	Since 15 years

We also asked users to provide us with some technical information about their experience in dealing with operating systems, screen readers, and what are the main problems when they navigate the Web. Table 3 shows a summary of answers for these technical questions. The first two columns indicate the number of operating systems (either Windows or Linux) used either on fixed or portable computers. The third, fourth, and fifth columns indicate the number of users who use JAWS (Job Access With Speech), NVDA (NonVisual Desktop Access), and ORCA, either on fixed or portable computers.

Table 3. Used operating systems and screen readers

	Windows	Linux	JAWS	NVDA	ORCA
Number of users with fixed computer	4	0	4	2	0
Number of users with portable computer	2	1	2	1	1

No one of the five blind persons uses a tablet, and the screen readers used with cellular phones are Talks and MobileSpeak with Nokia, and Voiceover with IPhone. Only one of the 5 persons uses a telephone to access the Web (access via IPhone). The main problems of accessing the Web via fixed or portable computers, or via IPhone telephone were: problems of access to Flash files, problems of AJAX technologies, and no ability to know the global structure of web sites. (These problems have been reported to us by the 5 users).

5.1.2 Explaining the objective of the test

To give the blind persons a good idea about the test, we explained in details what are the objectives and the phases of each tasks, and described contents of the embedded system; we also explained the final objective of the project, and why we concentrate on vibro-tactile technique regardless of other techniques.

This phase was important to initiate users for accepting kindly the test and for doing their best to interact with next steps as correctly as possible.

5.1.3 Training and learning task

In this training task, the user discovers the graphical elements in each image presented in figure 4 (images a, b (NT2), c (NT3), and d (NTG)), and the users were informed of each shape name. This task was very important for users to test the system before the

evaluation task, and to know exactly how the system transforms different the grey level under the touched points on the tablet screen to a vibration mode. It is also very useful for users to control their speed of mapping the screen either to discover either the borders or the contents. During this task, the program recorded the touching information in log files (X, Y coordinates, Pressure, and Time).
Table 4 indicates training times in minutes for each user, and for each image in figure4.

Table 4. Times of training task for each user (in minutes)

User ID / Image	ID0	ID1	ID2	ID3	ID4	Total	Average
A	4,99	2,60	3,55	9,97	3,96	25,08	5,02
b (NT2)	4,40	3,11	0,99	3,30	1,00	12,80	2,56
c (NT3)	2,81	6,29	2,54	2,85	1,15	15,65	3,13
d (NTG)	3,11	3,98	2,02	3,22	1,32	13,65	2,73
Total	15,31	15,99	9,10	19,34	7,43		

We notice from table 4 that discovering the first image takes more time, and it is normal because it is the first experiment for blind users on this prototype. We notice also that there is a significant decrease in time between discovering the first and the last image in the training task. This could be an indicator that training users could decrease the time for discovering graphical elements. We can also notice the significant difference between different tested persons, for example user with ID4 needed 7.43 minutes to scan the images (A, NT2, NT3, and NTG), but the user with ID3 needed 19.34 minutes to scan the same images.

5.1.4 Evaluation task

In this task, firstly we asked each user to discover the image 5.a and to find how many squares inside it and to name each founded square, then we asked them to discover images 5.b (IDP1), 5.c (IDP2), 5.d (IDP3), and to describe them to us, and to redraw discovered shapes. Table 5 illustrates an evaluation of answers for the first question to name squares in image 5.a (The blind users have been asked to name the shapes in image 5.a, and for each shape, we evaluated the number of correct and incorrect answers). (In tables 5 and 6, the symbol ✓ represents a correct answer for the touched shape, and the symbol X represents an incorrect answer or inability to select the name of the touched shape).

Table 5. Results of questions for squares in image 4.a

User-ID/ Square Name	ID 0	ID 1	ID 2	ID 3	ID 4	Number of errors
NT2	✓	✓	✓	X	✓	1
NT3	✓	✓	X	X	✓	2
NTG	✓	X	✓	X	✓	2

We notice from table 5 that the lowest number of errors is assigned to image NT2, and it is the same result which we obtained during tests with sighted persons.

Results of answers for other questions related to images IDP1, IDP2, and IDP3 are summarized in table 6.

Table 6. Answers of questions for images IDP1, IDP2, IDP3

User-ID	ID0			ID1			ID2			ID3			ID4		
IDP	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Answers about number of rectangles	✓	✓	X	✓	X	X	✓	X	✓	X	✓	✓	✓	✓	✓
Answers about sizes of rectangles	X	X	X	X	X	X	✓	X	X	X	✓	✓	✓	✓	✓

We notice from data in tables 4, 5, and 6, that the best performance is for the user with ID4, and this may be because that this female user is the youngest between others, and it could be because that she was the only one that has already used touched devices (an IPhone in her case working with VoiceOver).

After answering questions about each image of images (IDP1, IDP2, IDP3), we asked each user to redraw the graphical elements founded in each touched image. Figure 6 views the redrawing results of the user ID4 (ID4 is the female user who gave best answers).

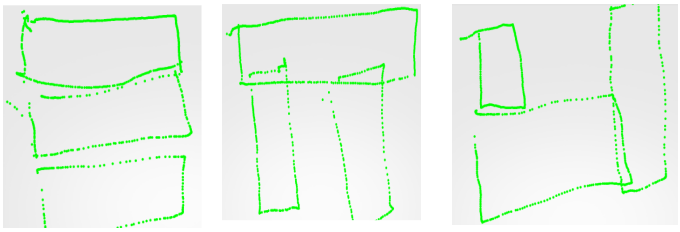


Figure 6. Results of redrawing images IDP1, IDP2, IDP3 for the user ID4.

When comparing results of redrawing (Figure 6) with images IDP1, IDP2, and IDP3, we find that the results are interesting, and we can conclude the following:

1. An ability of distinguishing sizes of shapes, because the degree of scaling between redrawn shapes is nearly equal to the degree of scaling between real shapes (IDP1, IDP2, IDP3).
2. An ability of distinguishing relations of directions, because relations of directions (vertical order, left to, right to,…) between redrawn shapes is nearly equal to relations of directions between real shapes.

The average of times in minutes consumed for each evaluation question is summarized in table 7.

Table 7. Times of the evaluation task for each user (in minutes)

User ID / Image	ID0	ID1	ID2	ID3	ID4	Total	Average
A	1,23	9,87	5,39	7,84	1,85	26,17	5,23
IDP1	4,39	14,99	1,41	3,75	1,41	25,96	5,19
IDP2	7,70	9,22	0,79	1,99	13,85	33,55	6,71
IDP3	2,71	12,94	2,81	4,03	12,58	35,06	7,01
Total	16,03	47,02	10,39	17,61	29,68		

5.2 Results Analysis

To get an idea about the most touched and the least touched areas on the screen during learning and evaluation tasks, we divided the touched-screen into 16 areas (as in figure 7, r00...r03, r10...r13, r20...r23, r30...r33), and calculated the average of touches in each area for all users.

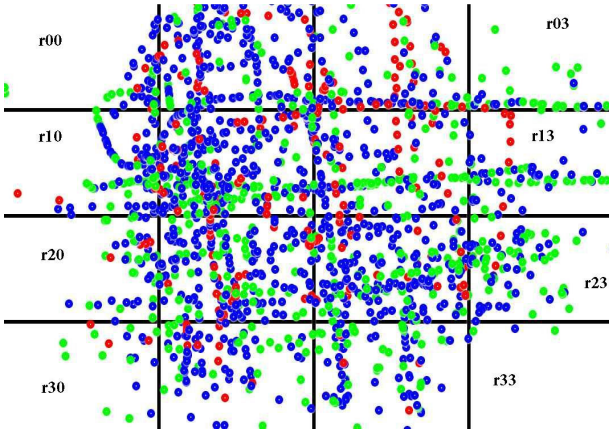


Figure 7. The 16 areas of the touched-screen of user with ID4 (Red points represent touched points with max pressure values, blue points represent touched points with pressure values less than the max and greater than the average, green points represent touched points with pressure values less than the average).

We have founded that the most touched areas are r12, r11, r22, r21, and the least touched areas are r30, r33, r32, r00 as described in figure 8. This information could be useful in next phases of our research in completing the graphical vibro-tactile language by putting the important information in the most touched areas.

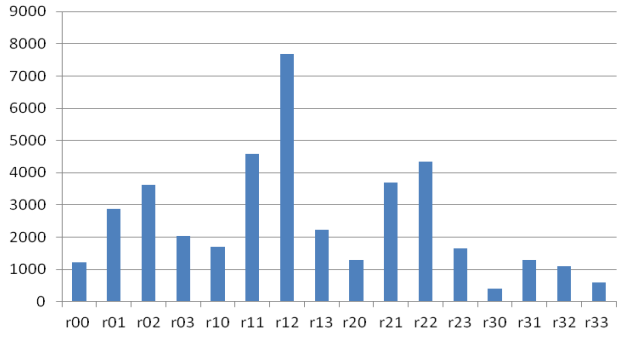


Figure 8. Most and least touched areas on the touched-screen.

During analysis the results, we have noticed that there are a lot of differences between the pressure values for all users (Pressure value depends on the used tablet; in these experiments we have used Asus Model TF101 with Android operating system).

To analyze pressure values, we calculated the max pressure value between all users, it was 3.19, and the average was 1.73, then we redrew the touched points for each user with considering that points with pressure values equal to the max value have been drawn in red color, points with pressure values less than the max and greater than the average have been drawn in blue color, and points with pressure values less than the average have been drawn in green color. Figure 7 represents an example of these points in different colors (for IDP3 evaluation task of the user who has id 4). After analyzing all the images drawn for all users, we have noticed that majority of red points (max pressure) are in images for which users gave right answers. This notice may be useful in designing our graphical vibro-tactile language, since we can observe when the user decreases or increases his touch pressure. The increasing of pressure may indicate that the user touches graphical elements interesting for him, and the decreasing may indicate that the user touches graphical elements non-interesting for him.

During the tests we observed also that users try sometimes to scan the screen very quickly, it might be because they try to get a lot of information in a short time.

6. CONCLUSION AND PERSPECTIVES

In this paper, we summarized our current work which aims to design an approach for non-visual access to web pages on touch-screen devices. The designed vibro-tactile protocol transforms the information viewed on the screen and touched by users to vibration by transforming light contrasts of touched pixels into low-frequencies tactile vibrations.

The obtained results are interesting, since we used in these experiments only one vibration motor of low level quality (phone vibrator), and the learning period was very short, so there are many enhancements to be achieved in next versions either on the hardware/software level or on the level of learning phase (increasing the number and quality of micro-vibrators, making more control on frequencies and amplitudes sent to micro-vibrators, adding vocal abilities to the current approach, integrating automatic intelligent methods for segmenting web pages,…).

In the same way that the environment enables a blind person to move in space with sidewalks and textures which will be explored by his/her white cane, we hope giving the blind user an ability to navigate documents depending on "textual sidewalks" and "graphical paths" which will be discovered by his/her finger.

Next steps in this research will be 1) Adding elements to the graphical vibro-tactile language in order to represent more HTML elements such links, buttons, input fields, and other elements, 2) Making the program more interactive to guarantee vibrating in real time, and without any delay. 3) Including the results obtained in these pre-tests, for example focusing on areas which are most touched by the users, focusing on the variation of pressure for expecting the model of navigation, 4) adding more vibration motors to the designed circuit; current version includes 2 vibration motors, but we tested only one to know exactly how blind users navigate the screen using one finger; next tests may be on more vibration motors to discover how the blind users navigate the screen using more than one finger, 5) We plan also to add thermic actuators for translating the notion of colors. This may be very useful and hopeful for blind users to transfer information about colors, 6) after adding all mentioned desired changes to the circuit and to the program; we should test integrating the hybrid segmentation algorithm of web pages with the adapted version of the designed circuit to generate automatically the graphical elements of the navigated web page.

Acknowledgment

This works is founded by the national agency of research ANR (Agence Nationale de Recherche http://www.agence-nationale-

recherche.fr/) as a part of ART-ADN project ANR-12-SOIN-0003-02 (Accès par Retour Tactilo-oral Aux Documents Numériques) in GREYC laboratory (www.greyc.fr) of the University of Caen Basse-Normandie (www.unicaen.fr) and the National Superior Engineering School of Caen (www.ensicaen.fr).

Thanks to Rabeb BEN SASSI, and Abdelmajid TOUNSI for their participation in designing the system.

7. REFERENCES

[1] Maurel, F., Dias, G., Routoure, J-M., Vautier, M., Beust, P., Molina, M., Sann, C., 2012. *Haptic Perception of Document Structure for Visually Impaired People on Handled Devices*, Procedia Computer Science, Volume 14, 2012, Pages 319-329, ISSN : 1877-0509.
DOI=http://dx.doi.org/10.1016/j.procs.2012.10.036

[2] Maurel, F., Vigouroux, N., Raynal, M., Oriola, B., 2003. *Contribution of the Transmodality Concept to Improve Web Accessibility*. In Assistive Technology Research Series, Volume 12, 2003, Pages 186-193. International conference; 1st, Smart homes and health telematics; Independent living for persons with disabilities and elderly people. ISSN : 1383-813X.

[3] http://www.chromevox.com/ [Access 7/5/2014]

[4] http://www.synapseadaptive.com/gw/wineyes.htm [Access 7/5/2014]

[5] http://www.freedomscientific.com/ [Access 7/5/2014]

[6]https://play.google.com/store/apps/details?id=es.codefactory.android.app.ma.vocalizerfrfdemo&hl=fr [Access 7/5/2014]

[7]https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback&hl=fr [Access 7/5/2014]

[8] http://www.apple.com/fr/accessibility/ [Access 7/5/2014]

[9] Alaeldin, A., Mustafa , Y., Sharief, B., 2012. *Tactile WebNavigator Device for Blind and Visually Impaired People*. In Proceedings of the 2011 Jordan Conference on Applied Electrical Engineering and Computing Technologies, Jordan, 2012.

DOI=http://dx.doi.org/10.1109/AEECT.2011.6132519

[10] BouIssa , Y., Mojahid, M., Oriola, B., Vigouroux, N., 2009. *Accessibility for the Blind, an Automated Audio/Tactile Description of Pictures in Digital Documents*. IEEE International Conference on Advances in Computational Tools for Engineering Applications, 2009, Pages: 591 – 594.

DOI=http://dx.doi.org/10.1109/ACTEA.2009.5227855

[11] Lenay, C., Gapenne, O., Hanneton, S., Marque, C., Genouëlle, C., *Sensory Substitution, Limits and Perspectives*. In Touch for Knowing Cognitive psychology of haptic manual perception, Amsterdam, 2003, Pages: 275-292.

[12] Tixier, M., Lenay, C., Le-Bihan, G., Gapenne, O., Aubert, D., *Designing Interactive Content with Blind Users for a Perceptual Supplementation System,* TEI 2013, 2013, in Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction, Barcelona, Spain, 2013, Pages 229-236.
DOI= http://dx.doi.org/10.1145/2460625.2460663

[13] MAUREL, F., SAFI, W., BEUST, P., ROUTOURE, J.M., 2013. *Navigation aveugle sur dispositifs mobiles : toucher le Web... pour mieux l'entendre*, 16ème Colloque International sur le Document Électronique, CIDE16, Lille France, Europia productions, 2013.

[14] Rotard, M., Knödler, S., Ertl, T., 2005. *A Tactile Web Browser for the Visually Disabled*. In Proceedings of the sixteenth ACM Conference on Hypertext and Hypermedia. ACM, New York, NY, USA, 2005, pages 15-22.

DOI= http://dx.doi.org/10.1145/1083356.1083361

Constructing Narrative Visualizations as a means of Increasing Learner Engagement

Bilal Yousuf
KDEG, Trinity College Dublin
Dublin, Ireland
yousufbi@scss.tcd.ie

Owen Conlan
KDEG, Trinity College Dublin
Dublin, Ireland
Owen.Conlan@scss.tcd.ie

ABSTRACT
Increasingly visualization systems are using storytelling to present complex data. However, many approaches neglect enabling users to independently explore details within the story. The research presented in this paper provides an overview of the implementation and discusses the evaluation of a novel framework (VisEN), which aims to allow users to construct narratives containing multiple exploration paths. The narratives are told through dynamically generated visualization techniques, which are personalized for individual end users, and where every visualization technique in the narrative can be further explored. The evaluation described assesses the role personalized visual narratives had in increasing engagement of weaker students with an online database SQL course. It was found that weaker students who regularly interacted with their personalized visual narratives showed an improvement in engagement.

Categories and Subject Descriptors
H.3.5 [Online Information Services]: Web-based services; H.5.2 [User Interface] Graphical User Interface; H.5.4 [Hypertext/ Hypermedia]: Architectures

General Terms
Design, Experimentation, Human Factors, Performance

Keywords
Visualizations, Personalized Visual Narratives, Visual Interaction and Exploration

1. INTRODUCTION
Research in the field on Information Visualization has largely been focused on visual analytics and exploration, whereas research in visual presentation and storytelling has recently started to gain momentum. Storytelling in information visualization, or narrative as it is referred to in this work, can be defined as an ordered sequence of steps consisting of visualizations, which are linked or connected to make the communicated message more memorable [1]. Stories provide effective ways of highlighting facts, making points and passing on information [16], while visualizations facilitate a simple means to understand digitized data as they map data attributes to visual properties [6]. The research addressed in this paper presents a framework, VisEN (Visual Exploration with Narrative), which aims to provide a novel way to extract knowledge and meaning from data. VisEN supports users in the role of narrative composers to analyze potentially complex data through advanced web based interfaces to construct narratives. The narratives include explorations paths to facilitate data drill downs and viewing related data. The narratives are automatically transformed into personalized visual narratives for end users, who can analyze and explore sections of the narrative through multiple interactive visualization techniques and gain a deep understanding of the data.

This paper discusses the implementation overview, evaluation and preliminary results of two key components of the VisEN framework: the Narrative Builder and the Visual Narrative Explorer. The aim of the Narrative Builder is to enable narrative composers to construct explorable narratives through an advanced web–based interface, which enables the analysis of potentially complex data without dealing with data complexity issues. The aim of the Visual Narrative Explorer is to personalize the visual narratives for end users and facilitate analysis and exploration of these narratives. VisEN was deployed to the AMAS [20] Personalized Learning Environment (PLE), to provide personalized visual narratives to 108 students who participated in an online SQL course. Two evaluations were completed with the first analyzing how effective the AMAS course professor found the user interfaces provided by the Narrative Builder to build explorable visual narratives. The second evaluation focused on weaker students' level of engagement ("participation in educationally effective practices" [17]). In particular, it analyzed how effective the personalized visual narratives were in allowing weaker students to extract meaning from their activity data, in order to motivate them to engage with the course. The results of both evaluations were very encouraging and it was found that these learners were drawn to their visual narratives in order to understand and improve their engagement with the course.

The remainder of this paper is structured as follows: Section 2 discusses the VisEN framework approach. Section 3 presents a review of the related work. Section 4 describes an implementation overview of VisEN. Section 5 presents two use cases; the first describing a domain expert using VisEN to construct visual narratives, and the second describing a learner using her personalized visual narratives to gain a thorough understanding of her personal course log data. Section 6 evaluates effectiveness of VisEN when deployed to a PLE and discusses preliminary results. Finally, section 7 discusses conclusions and future work.

2. VISEN APPROACH
VisEN automatically transforms narratives into explorable visual narratives. This transformation requires data characterization and

mappings to transform data to appropriate visualization techniques. Data characterization or data transformation [6] involves analyzing data to facilitate automated mappings to visualization techniques. To enable this mapping or visual encoding [6], the affordances and characteristics of visualization techniques are required, for example, through a matrix. VisEN narratives consist of data slices, which are constructed using data fields, metadata, filters and aggregations. Data slices form the chapters or sections of the narrative.

When a data slice is constructed, visualizations that can render the data are automatically generated and presented to the narrative composer as a set. The narrative composer decides which visualizations to keep in the set. This action introduces humans into the visual matching process. This results in a refined set of visualizations for a data slice, and takes place before the narrative is transformed into a visual narrative. VisEN automatically generates personalized exploration paths to allow end users to select elements within visualizations and view details or view related data through other visualizations. The exploration paths are generated based on users preferences and consists of visualizations showing details and related data to the narrative viewed.

To complete the narrative, the narrative composer connects the data slices to each other in a chronological order and publishes it. Figure 1 shows a simplified view of the process used by VisEN to produce personalized explorable visual narratives.

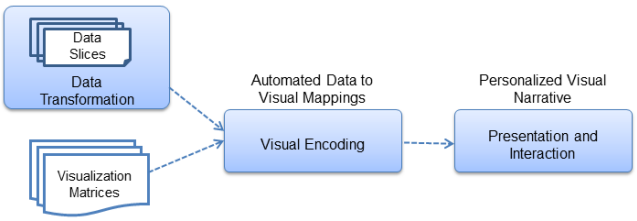


Figure 1: VisEN Flow

3. RELATED WORK

Interaction, exploration and visual storytelling are important aspects of presentation in information visualization as they allow users to gain a deeper understanding of data. This section analyses the state of the art to determine how adequately generating dynamic visual narratives and enabling personalized visual explorations of these narratives have been addressed.

Visual narratives have been effectively used in journalism [9, 15, 24] to tell stories with data. These have ranged from presenting several visualizations with annotations in one view to slides containing interactive visualizations to tell a story. Contextifier [15] for example, provides visualizations embedded in news articles and provides visualizations of related articles allowing users to navigate and explore these. Tools such as Gapminder [22], GED Viz [8] and SketchStory [18] provide users with interactive visual storytelling. However, the interactions are limited to hovering the mouse over data points to reveal details and filtering regions of the data. StoryFlow [19] allows users to explore data in a second layer of the story through its bundling operation, which reveals a level of detail beneath a bundled line. Spotfire [27] provides users with data drill down capabilities, where visual structures can be clicked by users and the system loads another visualization that also provides a drill down of the data. A user can choose to drill down further and view the selected data through a further visualization. However, with drill downs, users reach an end point where their exploration must. Exploration paths provided by VisEN are linked to elements in the visual narrative and when these elements are clicked, visualization techniques are generated rendering a drill down view or a related data view of the element selected. Drill down views show the details surrounding a selected element, whereas related data views show data which shares relationships with the selected element. When a user reaches the lowest point in a drill down, she always has the option to view related data. Visualizations have been used in Technology Enhanced Learning (TEL) to present student activity data and peer comparisons [11, 22] to motivate students. However, these are not represented through visual narratives, where users can explore the data presented.

Personalized visual narratives can aid the process of understanding complex data as they can present personalized data and provide visualizations that suit individual preferences. In Tableau Story, Tableau [26] selects the most suitable visualization for the story point and this can be changed by the analyst. Similarly Google Fusion Tables [10] uses a suitable visualization for the data. However, we find on many occasions, a number of visualization techniques are suitable to render the same data. The visualizations generated by these systems are not personalized to end user preferences. In TEL, a number of systems [2, 3, 21] provide personalized visualization forming part of the learning module. VisEN's architecture consists of a Personalization Engine, which generates personalized exploration paths for end users. User data preferences are stored in a user model, which are used to personalize the exploration paths.

From the visualization tools that support visual interactions and explorations, Spotfire [27] supports drill down explorations, however, the exploration path is fixed and an end user has the option to either view the details behind a data point or not. The exploration is not independent of the path constructed by the analyst. VisEN provides multiple exploration paths from each data slice, allowing end users to explore various tailored paths through the data set. Hence the exploration is independent from one end user to another and this allows users to derive personal conclusions.

From the analysis above, it can be seen that VisEN progresses the state-of-the-art by introducing three novel factors which focus on allowing end users to: 1) explore related data through exploration paths; 2) view visual narratives; and 3) analyze tailored exploration paths.

4. IMPLEMENTATION OVERVIEW

The VisEN architecture uses principles discussed in 1) the visualization pipeline [6]; 2) the visual information seeking mantra [25]; 3) the Template Editor and Shelf Configuration visual interface design approaches [13]; and 4) sequencing in visual narratives [14] to generate explorable personalized visual narratives. Figure 2 shows VisEN architecture, which consists of the Narrative Builder, the Visualization Engine and the Visual Narrative Explorer components.

4.1 Narrative Builder

The Narrative Builder enables narrative composers to easily construct narratives from complex data. Visualizations are not introduced into the narrative during the narrative building phase.

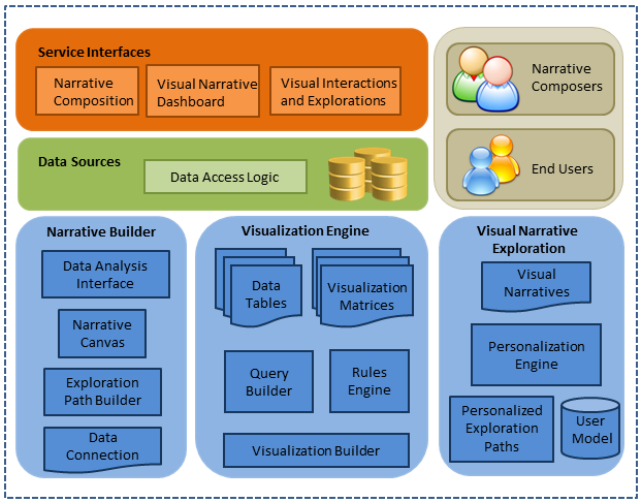


Figure 2. VisEN Architecture

4.1.1 Data Connection Component

Narrative Composers use the Data Connection component to connect to heterogeneous data sources to construct narratives. Data connections are established by selecting data sources or specifying connection parameters. Preconfigured data source parameters are stored in configurations files and new data source parameters supplied by narrative composers are also saved to these files.

4.1.2 Data Analysis Interface

Data slices form the individual pieces of narratives and are constructed by the narrative composers via the web based Data Analysis Interface. In addition to constructing the data slice, the Data Analysis Interface allows narrative composers to analyze data sources. The interface consists of a number of buttons which run general queries such as "*select count..*", "*select <field>..*" etc.; this simplifies the process of constructing narratives as the raw data values can be analyzed by narrative composers. The Data Analysis Interface uses the jQuery Accordion widget to show source tables and fields and uses the jQuery Draggable widget to facilitate dragging and dropping of data fields to construct data slices. The interface provides a canvas with panels for fields and filters. The data fields from the Draggable widget can be dropped onto these panels to construct data slices. The drag and drop design approach has been used effectively in state of the art [26]. When a field is dropped onto a filter panel, VisEN runs queries to fetch data to allow narrative composers to specify which values to use in the filter.

4.1.3 Encoded Exploration

An important and novel aspect of VisEN is exploration paths, which are automatically constructed and connected to data slices. Exploration paths consist of a series of visualizations linked to each data slice or section of the narrative. End users can view and analyze exploration paths by clicking on elements in a data slice to drill down into sections of a narrative or explore related items to obtain a deeper understanding of the data. Exploration paths are constructed by VisEN using data slices that have common elements or derivatives in the narrative. The narrative composer can view the automatically constructed exploration paths and can remove and visualization to the path via the available add/remove options on the Data Analysis Interface.

4.2 Visualization Engine

The Visualization Engine transforms narratives into visual narratives by mapping data slices from the narrative to visualization techniques.

4.2.1 Query Builder

The Query Builder uses the data and metadata provided by the narratives composers in the data slices to generate and execute SQL queries against the specified data sources. The query results are formatted by data type, size (data sizes and number of series of data) and coordinates (data points) to aid the Rules Engine in selecting appropriate visualizations for the data slice.

4.2.2 Rules Engine

The Rules Engine uses the formatted query results and the data slice metadata to determine appropriate visualization techniques for each data slice of the narrative. Instead of building visualizations, VisEN utilizes JavaScript visualization libraries to source visualization techniques. Extensive research [4, 5, 7, 12] has evaluated the affordances and characteristics of visualization techniques and compared the suitability of various techniques for data sets. This research has been used by VisEN to allow developers to build matrices that specify the characteristics, affordances and constraints of the supported visualizations. The matrices are stored as XML files and new visualizations can be seamlessly incorporated into the framework by creating a new XML file (matrix) and importing the JavaScript library.

4.2.3 Visualization Builder

The current set of visualization techniques supported by VisEN requires data to be formatted as JSON objects. The Visualization Builder creates JSON objects using the query results and metadata and populates the set of visualization techniques (currently nine techniques are supported including: bar chart, bubble chart, gauge, line chart, pie chart, scatterplot, stacked bar chart, area chart and parallel coordinates). It also makes the populated set of visualizations available to the narrative composer to view through a web interface as a dropdown list, where visualizations can be removed from the set. The remaining set is used for the visual narrative.

4.3 Visual Narrative Explorer

The Visual Narrative Explorer personalizes the visual narratives for end users by generating tailored exploration paths for each narrative based on individual preferences. It provides a web-based interface where end users can analyze visual narratives and view exploration paths to understand data.

4.3.1 Personalization Engine

The Encoded Exploration component generates derivatives from data slices for exploration paths, which can be accepted or rejected by the narrative composer. Accepted derivate data slices and data slices related to the narrative are used to form personalized exploration paths. The Personalization Engine personalizes the exploration paths using user data preferences, set in the user model. These preferences are set when end users asked to select data tags (taken from data slice metadata) they are interested in exploring when viewing visual narratives. Selected tags are stored in the VisEN user model and these are used to personalize the exploration path.

4.3.2 Narrative Dashboard

Published visual narratives are made available to end users through the web based Narrative Dashboard. End users are presented with the first data slice of visual narratives and the remaining data slice can be access by clicking the titles at top of the interface. When an end user wishes to explore an element in the data slice, she can click it and this generates the first visualization in exploration path, which is shown in a popup window on the web browser. Clicking an element in the visual narrative fires an AJAX request and the linked exploration path is made available to the end user. At any point the end users can close the exploration path popup window and continue analyzing the visual narrative or alternatively continue with the exploration.

5. USE CASES

This section discusses two use cases; the first use case describes a university professor using VisEN to construct two narratives. The second use case describes a student using personalized visual narratives to understand and improve her course engagement.

5.1 Use Case One – University Professor

John is a Professor lecturing Database Management System to final year university students. His students need to use the AMAS [20] portal to study SQL. John understands the challenges learners' have engaging with online learning modules and wishes to provide visual narratives to improve engagement by allowing them to visually analyze and explore their individual log data.

John logs into VisEN and assumes the role of a narrative composer. He connects to the AMAS data source containing learner log data from the last time the course was run. This data source consists of thousands of entries with all the interactions learners had with the course over a three months period. After analyzing the data he wishes to construct two narratives. He starts constructing data slices by dragging data fields onto the Narrative Builder interface.. He clicks on the "Visualize Data" button and views the set of visualizations for each data slice and also views the automatically generated exploration paths. Finally he disassociates the narrative with the previous log data and connects it to new data source (this consists of test entries as the course is yet to commence) and publishes the narratives.

5.2 Use Case Two – Final Year Student

Michelle is a final year Computer Science student and has received an average grade of below 50% each year during the first three years of her course. However, she is determined to improve her grade in her final year. As part of one of her modules she needs to study SQL using the AMAS portal. During the first month of the three month module, Michelle has occasionally used the portal. At the end of this month she receives a notification from the portal informing her of her poor engagement with course activities and advises her that in previous years the students who continued to engage at this level performed poorly.

Following on from this notification, Michelle wants to understand how she can improve her engagement and estimate how much time she must commit to this module to perform well. She views her personalized visual narratives and analyzes her engagement score and how it was calculated. She analyzes peer engagement comparisons using her visual narratives which allow her to determine how to improve engagement. By analyzing peer comparisons and exploring her visual narratives, Michelle is able to predict how long it will take her to complete her next five activities. Michelle now feels motivated and determined to work hard and obtain a good grade. As she completes each activity, she explores her visual narratives and estimates the time the next activity would take.

6. EVALUATION

VisEN was deployed to the AMAS [20] PLE during the 2013-2014 academic year to provide learners with personalized visual narratives to allow them to analyze their engagement score, view time spent on activities and analyze peer comparisons. AMAS provides a dynamic and adaptive framework for composition and assignment of personalized learning activities [20]. It has been used over the past three years to deliver an SQL database course to final year university students in Trinity College Dublin. Two evaluations were carried out in conjunction with the delivery of the AMAS SQL course. The first evaluation involved a university professor using VisEN to construct visual narratives for his students. The second evaluation involved participating students of the course using personalized visual narratives in order to understand their performance and engagement from their log data.

6.1 Evaluating the Narrative Builder

In this evaluation, the professor whose students worked through the AMAS activities, assumed the role of a narrative composer and constructed narratives using the AMAS log data from the 2012-2013 academic year. The aim of this trial was to evaluate the end to end tasks of the narrative composer: analyze a complex data set; construct narratives with exploration paths; and critique the set of generated visualizations. The professor was provided with a 15 minutes training session on how to use the Narrative Builder and then asked to construct the two narratives using the Narrative Builder (shown on the left of figure 3): 1) A narrative showing learners' engagement score and how it was calculated; 2) A narrative presenting the time learners spent on activities, and allowing learners to compare activity times with their peers. Exploration paths were automatically generated, which showed a breakdown of selected students' engagement score (drill down). The other exploration path showed engagement scores of similar students (related data). Once both narratives were completed (which took 25 minutes with some assistance), the professor was asked to interact with the visual narratives, which were automatically generated and analyze the data through exploration paths. During the analysis, he was asked to answer questions by exploring and interacting with the visual narratives, which he did with ease and answered all the questions.

His final task was to critique the visualizations and the process of constructing the narrative through a questionnaire and interview. The questionnaire focused on how useful the professor found the process of constructing narratives and analyzing exploration paths. For example, one of the questions asked: "When viewing course engagement by activity, how useful was it to view students with similar engagement through an exploration path". The questions also addressed how well the framework and visualizations met his needs, such as "Did the framework support you in telling the story you wanted to tell" and "Where you ever frustrated with the limitations of the user interface", to which he offered useful suggestions such as providing tooltips and help options. From the feedback the professor found exploration paths very useful for gaining insight and was able to tell the story

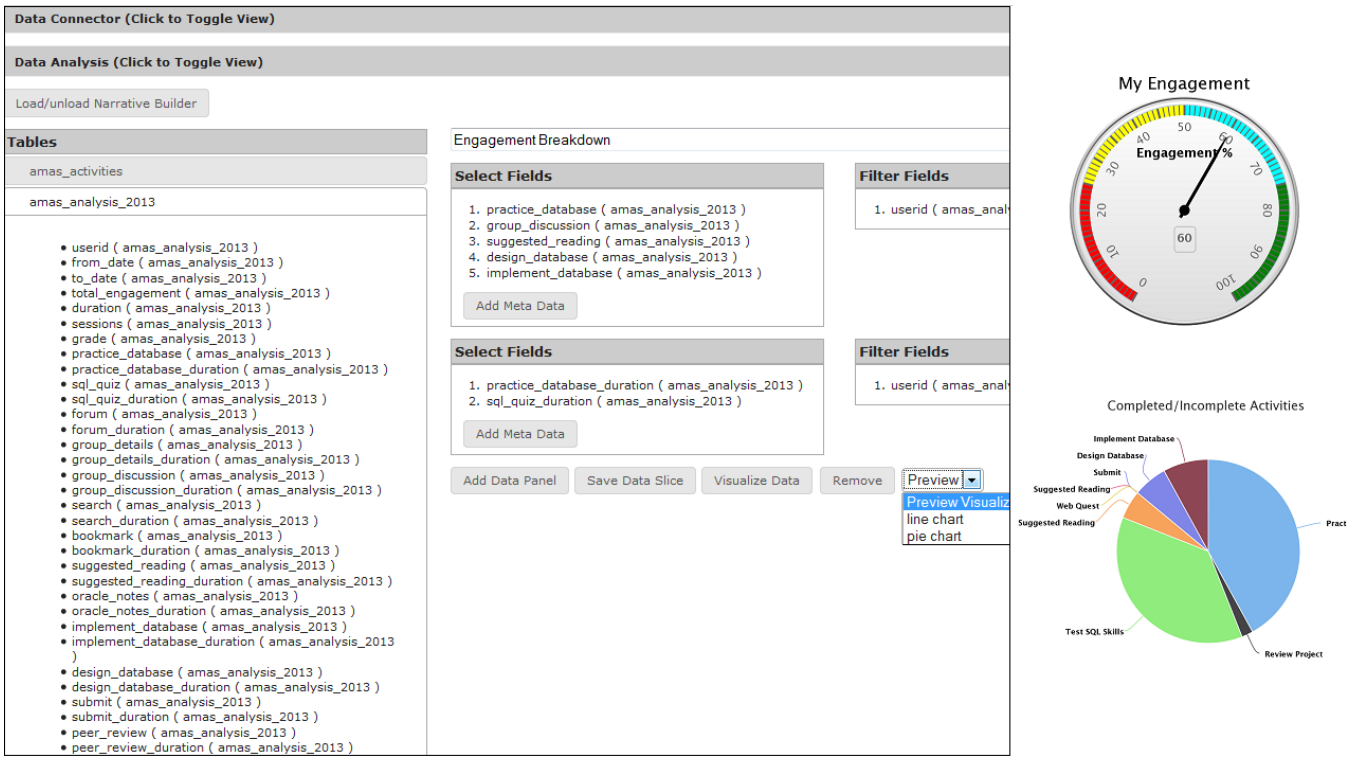


Figure 3. Narrative Builder Interface (left) and two sample visualizations from a Personalized Student Visual Narrative (Right)

requested. He expressed that the data slices and resulting visualizations represented his needs quite well. In the interview, the professor expressed that he was able to follow and interact with the visualizations easily and expressed confidence in constructing data slices and building the narratives. Examining the time taken to learn and construct the narratives, it was evident that the professor had a very positive experience constructing narratives using the Narrative Builder.

6.2 Evaluating Personalized Visual Narratives

One of the primary aims of AMAS is to support weaker students completing their course. The second evaluation focused on analyzing the impact the personalized visual narratives had on supporting weaker learners to improve course engagement. The right hand side of Figure 3 shows two visualizations from one of the narratives presented to learners. 108 students participated in the AMAS SQL course; 22 of these were identified as weak students as they had an average grade of below 50% for each of the previous three years of their course.

During the course, AMAS sent fortnightly notifications to learners informing them of their engagement levels. The first study analyzed the AMAS log data, (consisting of thousands of entries for three months of interactions from 108 learners), and found that all of the weaker students had at some stage received a below average engagement notification. The analysis of the log data of the 22 weaker students found that 17 of these students showed an improvement in engagement following this notification. It was found that 14 of these 17 learners were immediately drawn to their personalized visual narrative following a below average engagement notification. All of these 14 learners executed a minimum of 45% of their total narrative interactions on the first day after reading the notification. Following this notification (which did not explicitly direct them to their personalized visual narratives), these learners frequently returned to view their personalized visual narratives. Hence, it can be concluded that the personalized visual narratives assisted these learners in gaining a deeper knowledge of their performance data.

The second study analyzed if there was a correlation between weaker students interacting with their visual narratives and an improvement in engagement. The log data of the 17 weaker students, who showed engagement improvement following a below average engagement notification, was analyzed. It was found that all of these learners showed a minimum of a 70% increase in interactions with their visual narratives during the period in which their engagement improved. From this, it was concluded that weaker students who increased in interactions with their personalized visual narratives showed an improvement in their course engagement level.

7. CONCLUSIONS AND FUTURE WORK

This paper introduced VisEN as a framework to construct visual narratives and facilitate personalized visual explorations by allowing end users to: 1) explore related data; 2) analyze visual narratives; and 3) analyze personalized exploration paths.

Two evaluations were carried out; the first evaluation involved a university professor analyzing the log data of his students' course activities and constructing visual narratives. The results of this evaluation were positive, with the professor confidently creating data slices and narratives and positively commenting on his experience of executing the tasks required. The second evaluation involved analyzing the log data of weaker students who participated in an online SQL course. This evaluation found that the personalized visual narratives assisted these learners in understanding and improving their engagement and performance data.

Preliminary results have been obtained from both evaluations. Further work is required to evaluate the Narrative Builder through qualitative and quantitative analysis using several users. In the 2014 - 2015 academic year, it is intended to continue to provide learners with personalized visual narratives and compare engagement results with control groups, and quantify the increase in engagement levels, and verify the statistical significance.

8. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College Dublin.

9. REFERENCES

[1] Austin, M. *Evolution, Anxiety, and the Origins of Literature*. University of Nebraska Press, 2011.

[2] Brusilovsky, P., Ahn, J. W., Dumitriu, T. and Yudelson, M. 2006. Adaptive knowledge-based visualization for accessing educational examples. In Proceedings of *Tenth International Conference on* Information Visualization (London, UK 5-7 July 2006). IEEE pp. 142-150. DOI= 10.1109/IV.2006.16

[3] Brusilovsky, P. and Loboda, T. D. 2006. WADEIn II: A case for adaptive explanatory visualization. *ACM SIGCSE Bulletin.* 38, 3, 48-52. ACM, New York, NY.

[4] Chi, E. H.2000. A taxonomy of visualization techniques using the data state reference model. In *Proceedings of the IEEE Symposium on Information Visualization* (Salt Lake City, Utah, USA, October 9-10 2000), IEEE, pp. 69–75. DOI= 10.1109/INFVIS.2000.885092

[5] Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. 1987. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association, 82*, 398 (1987), 424-436.

[6] Card S. K., Mackinlay J. D., Shneiderman B. 1999. *Readings in Information Visualization: Using Vision to Think*. Ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[7] Dias M. M. Yamaguchi J. K. Rabelo E. and Franco C. 2012. *Visualization techniques: Which is the most appropriate in the process of knowledge discovery in data base?* InTech (September 2012). DOI=10.5772/50163

[8] Esche A. 2013. Ged viz Retrieved, November 10, 2013 from http://viz.ged- project.de/

[9] Gao, T., Hullman, J., Adar, E., Hecht, B., and Diakopoulos, N. 2014. NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News. In *Proceedings of the Conference on Human Factors in Computing Systems* (Toronto, Canada April 26–May 1, 2014). *(CHI)* ACM, New York, 3005-3014. DOI= 10.1145/2556288.2557228

[10] Gonzalez H. et al. 2010. Google fusion tables: Data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM symposium on Cloud computing* (Indiana, USA, June 6–11, 2010). ACM, New York 175-180. DOI= 10.1145/1807128.1807158

[11] Govaerts S., Verbert K., and Duval E. 2011. Evaluating the student activity meter: Two case studies. In *Proceedings of the International conference on Advances in Web-Based Learning*. Springer, Dec. 2011, pp. 188-197.

[12] Graham M. and Kennedy J. 2010. A survey of multiple tree visualization. *Information Visualization* 9, 4 (Dec. 2010), 235–252. DOI= 10.1057/ivs.2009.29

[13] Grammel, L., Bennett, C., Tory, M., & Storey, M. A. 2013. A Survey of Visualization Construction User Interfaces. In *EuroVis-Short Papers* (pp. 19-23). The Eurographics Association.

[14] Hullman J[1], Drucker S, Henry Riche N, Lee B, Fisher D, Adar E. 2013. A deeper understanding of sequence in narrative visualization. *IEEE Trans Vis Comput Graph*. 19, 12 (Dec 2013), 2406-15. DOI= 10.1109/TVCG.2013.119.

[15] Hullman, J., Diakopoulos, N. and Adar, E. 2013. Contextifier: Automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France April 27-May 2). ACM, New York, NY 2707-2716. DOI= 10.1145/2470654.2481374

[16] Kosara R. and Mackinlay J. 2013. Storytelling: The next step for visualization. Computer 46, 5 (2013), 44–50.

[17] Kuh, G. D. 2007. How to Help Students Achieve. *Chronicle of Higher Education*. 53 (41), pp. B12–13.

[18] Lee, B., Kazi, R. H., and Smith, G. 2013. SketchStory: Telling more engaging stories with data through freeform sketching. *Visualization and Computer Graphics, IEEE Transactions on, 19*,12 (2013) 2416-2425.

[19] Liu, S., Wu, Y., Wei, E., Liu, M., & Liu, Y. (2013). Storyflow: Tracking the evolution of stories. *Visualization and Computer Graphics, IEEE Transactions on*, *19*(12), 2436-2445. DOI=10.1109/TVCG.2013.196

[20] O'Keeffe I. et al. 2012. Personalized activity based eLearning. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies* (Graz, Austria, September 5-7, 2012), i-KNOW '12, ACM,New York, NY, Article 2.

[21] Pierson, W. C., and Rodger, S. H. 1998. *Web-based animation of data structures using JAWAA*. In ACM SIGCSE Bulletin, 30, 1, (1998). 267-271. DOI= 10.1145/274790.274310

[22] Santos J., Govaerts S., Verbert K., and Duval E. 2012. Goal-oriented visualizations of activity tracking: A case study with engineering students. In *Proceedings of the International Conference on Learning Analytics and Knowledge*, ACM, May 2012, pp. 143-152, DOI=10.1145/2330601.2330639.

[23] Rosling H. 2013. Gap minder viz Retrieved, March 1, 2014 from http://gapminder.org.

[24] Segel, E., & Heer, J. 2010. Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on*, *16*,6, (2010) 1139-1148.

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

[26] Tableau software. 2013, Retrieved January 15, 2014 from http://tableausoftware.com

[27] Tibco spotfire.2013. Retrieved, October 5, 2013 from http://spotfire.tibco.com