UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

GUILHERME BERTINI BOETTCHER

Visualization of News and Press Information Diffusion in Social Media

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Carla Maria Dal Sasso Freitas
Coadvisor: João Comba

Porto Alegre
March 2016

CIP – CATALOGING-IN-PUBLICATION

Bertini Boettcher, Guilherme

Visualization of News and Press Information Diffusion in Social Media / Guilherme Bertini Boettcher. – Porto Alegre: PPGC da UFRGS, 2016.

68 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2016. Advisor: Carla Maria Dal Sasso Freitas; Coadvisor: João Comba.

1. Information visualization. 2. Social media. 3. Temporal patterns. 4. Diffusion patterns. I. Dal Sasso Freitas, Carla Maria. II. Comba, João . III. Título.

"A certain darkness is needed to see the stars."

— Osho

ACKNOWLEDGMENTS

Looking back and acknowledging all the people who were there to help me through the moments of doubt, the ones that made sure my path was true and gave so much while expecting so little and truthfully express the gratitude in my heart is one of the hardest tasks I have faced.

Mãe, pai, Edo e toda família, muito obrigado pelo apoio durante toda a minha existência, pela força para seguir em frente e pelo sempre presente amor e perseverança. Sem vocês, isso não seria nada mais que um sonho.

Carin, pelo exemplo e inspiração, pelo carinho e por me fazer enxergar através da neblina. Obrigado por me ajudar a sentir, obrigado por me ajudar a querer. Pelas incontáveis noites sozinha, pela paciência durante madrugadas sem dormir e por estar sempre pronta para me levantar, muito obrigado. Essa conquista é nossa!

To my advisor, Carla Freitas, for guiding me since my Bachelors, for believing in my ideas and giving me the freedom to pursue them and helping me accomplish so much and for the countless hours spent pointing me in the right direction over and over again.

To my co-advisor, João Comba, for being there every time, for believing in me and for being the friend I needed, when I needed it. Thank you for being truthful and for pushing me to get so far.

Special thanks to all the friends that challenged me to keep working and to run through the finish line.

Thanks to the Federal University of Rio Grande do Sul (UFRGS), the Institute of Informatics (INF) for the high quality education. I also thank my fellows in the Visualization, Interaction and Computer Graphics Lab who provided great company and for being a great source of inspiration during my research. Thanks to Luciana Nedel and Manuel Oliveira for their lessons, suggestions and help.

ABSTRACT

Social media has played a big part in the innovation process for newspapers and magazines, but adapting while going through a recession has led to a hasty evolution and automated processes for very different media. While existing social media studies and state-of-the-art visual solutions are available for analyzing social media content and users' behaviors, no other method is optimized for finding patterns from a popularity standpoint in the specialized realm of news channels. In this dissertation, we discuss two different usages of a combination of different visualization techniques that co-relate profiles' and their reading community activities with the resulting popularity.

We gathered Twitter posts, the number of followers and trending topics from worldwide press profiles as the data set foundation for our proposal. We used this data set as the seed for our visualizations to allow for multiple source comparison, so that not only the user is able to understand their own community but also the success and pitfalls faced by the competition in the same medium. We validate our analysis by interviewing a group of journalists from different established newspapers. Through interacting with our system, it was possible to detect hidden patterns in the massive dataset of messages and comments worldwide, enabling the users to have unique insights into their community's behaviors and preferences.

Keywords: Information visualization. social media. temporal patterns. diffusion patterns.

Visualização da difusão de informação de jornais e notícias em redes sociais

RESUMO

As mídias sociais tiveram grande impacto no processo de inovação de jornais e revistas, porém adaptar-se em momentos de recessão ocasionou uma evolução apressada e procedimentos automatizados para mídias essencialmente diferentes. Apesar de haver estudos e soluções estado-da-arte em visualização para midias sociais dedicadas à análise do conteúdo e do comportamento de usuários, nenhum método é otimizado (e especializado) para encontrar padrões do ponto de vista de popularidade no contexto de canais de notícias. Nesta dissertação, discutimos duas diferentes propostas de utilização de combinações de técnicas de visualização que correlacionam as atividades de perfis e de sua comunidade de leitores com a popularidade resultante.

Dados foram coletados reunindo postagens no Twitter, o número de seguidores dos perfis de interesse e os tópicos em tendência (*trending topics*) de canais da imprensa mundial para formar o conjunto de dados base para as propostas aqui apresentadas. Este conjunto de dados foi utilizado como semente para visualizações destinadas a possibilitar a comparação de múltiplas fontes, de forma que os usuários sejam habilitados a entender sua própria comunidade e também os sucessos e derrotas enfrentados pelos perfis competidores no mesmo meio. Essa análise foi validada através de entrevistas com um grupo de jornalistas de diferentes jornais estabelecidos. Através da interação com o sistema, é possível detectar padrões desconhecidos no grande volume de dados de mensagens e comentários de todo o mundo, possibilitando ao usuário uma visão única sobre os comportamentos e preferências de sua comunidade.

Palavras-chave: visualização de informação, midia social, processamento de imagens, padrões temporais, padrões de difusão.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

In the last few years, we have witnessed a dramatic change in the way newspapers and magazines communicate, as well as in the time events take to spread around the world through the Internet. The decline in sales of print media has forced the press to adapt its business to a more current media. One of the ways such media adapted was to create online profiles on social media such as Twitter and Facebook, and utilize them to create attraction to the articles posted on proprietary websites. In this way, news sources hope to reach the existing readers as well as increase their communities by allowing each reader to interact with the news pieces by commenting, favoriting and sharing them.

Facebook and Twitter, created as a social network and a micro-blogging service respectively, are popular social media venues that are recognized as relevant broadcasting and influence tools. Nowadays, they are commonly used by the press to generate interest for their published material and broadcast news. However, the adaptation from an established media to a new technology was hasty and for a variety of reasons, most newspapers simply utilize automatic publications to share their content on social media, with little to no research on their readers usage of such media.

The problem of identifying behavioral patterns in social media from a popularity standpoint is applicable in varied fields other than news. It can bring interesting insights for the fields of politics, entertainment or any other popularity driven networks that can benefit from finding key behavioral patterns in a social media. From the pool of motivating questions put together during our research, we selected the following as key:

- Why do readers stop following a profile? The most important aspect of a news source is their popularity, so it is extremely valuable to identify any pattern that results in a measurable change in the number of followers.

- Does the time of the post correlate with the number of re-tweets? It is important to find the best time to maximize attraction to an article in the community.

- How does the profile relate with its network? There are many analytics that can reveal the relationship between profile and reading community, such as the channels most used or re-tweeted words, the time that currently generates the most comments by the readers, and which sentiment is associated with the profile as whole.

- Is there a correlation between posted subjects and the amount of followers/popularity of a profile? Discovering patterns in the subjects of interest from the news source and the interest of readers can be very revealing regarding popularity and, therefore, the amount of attraction an online post can generate for a source.

Our research focused on the press community from a popularity standpoint. In the related literature, none of the solutions found offer a complete description that successfully answers all questions raised above. For example, the works by Guodao et al. (SUN et al., 2014), Yafeng et al. (LU et al., 2014), Yingcai et al. (WU et al., 2014) and Zhao et al. (ZHAO et al., 2014b) (ZHAO et al., 2014a) are the closest to a full solution, and we have studied their work prior to the design of our proposal.

The main objective of this work is to design a solution to visually demonstrate the behavioral patterns that answer our four key motivating questions. We propose the usage of a combination of different visualization techniques that correlate the profiles' and their reading community activities with the resulting popularity.

1.1 Structure of the Dissertation

We organize the remainder of the text as follows. In Chapter 2, we discuss the recent techniques, their advantages and shortcomings. In Chapter 3, we address the problem and its impact in different scenarios, and we also present the results obtained during the first phase of our research. In Chapter 4, we present our proposed solution pipeline in detail. We analyze the experimental results obtained in three different use cases and expert reports in Chapter 5, and finally, in Chapter 6, we provide a review of the technique, its advantages and shortcomings and final comments.

2 RELATED WORK

There are several works dealing with social networks analysis and visualization. Most are relatively simple tools to obtain statistics about users and the overall network. Literature closely related to our work can be roughly divided into two categories: analysis of information diffusion processes and visualization of diffusion patterns.

2.1 Analysis of Information Diffusion Processes

Social networks have been studied for years, but online social networks, blogs and microblogs introduced challenges in the investigation of how people communicate using these media. The analysis of information diffusion in such networks involve measuring quantitative characteristics, finding relations between structure and dynamics, predicting characteristics of the diffusion process and approaches for trends detection.

A complete overview of social networks can be found in (WASSERMAN; FAUST, 1994), where they present a review of network analysis methods and social network applications, with a focus on methods and models for analyzing social network data. They introduce the subject discussing its origins in behavioral sciences, then describe social network data in terms of structural and composition variables, modes and affiliation, boundaries and measurement. In the second part, the work discusses mathematical representation, structural and locational properties, roles and positions and finally ending the discussion with dyadic and triadic methods using example problems to elaborate the solutions.

In the realm of social networks quantitative characteristics, (KWAK et al., 2010) study the topological characteristics of Twitter and its power as a new medium of information sharing. They identify influential profiles by ranking users by the number of followers and by page rank, analyze the tweets of top trending topics and classify the trending topics based on the active period and the tweets and show that the majority of topics are headline news or persistent news in nature, concluding

that any re-tweeted post reaches an average of 1000 users, independently of the number of followers of the original author. Similarly, (YE; WU, 2010) present a measurement study of 58 million messages collected from 700000 users, analyze the propagation patterns and show how one case of breaking news spreads reaching similar conclusions to their counterpart.

The relationship between social network structure and dynamics, (LERMAN; GHOSH, 2010) compare data from Digg and Twitter and track how the interest in news stories spreads among them. They show how the networks impact the distribution of stories and that the network structure affects dynamics of information flow. (YANG; COUNTS, 2010a) similarly make a comparison of Twitter and a regular weblog network for their diffusion structures and find systematic differences between the two in contribution, navigation and interactive structural patterns, finding that microblogs have a unique role and characteristics in the social medial space, being more decentralized and connected locally, which indicates a flatter social structure and that some social network analysis algorithms like PageRank ((BRIN; PAGE, 1998)) are not applicable.

The characteristics of the diffusion process is discussed in (YANG; COUNTS, 2010b), where they focus on the mentions of authors for network analyses. They construct a model to capture the speed, scale and range of information diffusion. Finally, they compare the impact of posts properties versus the properties of the users, rate of user mentions historically, leaning on the latter for stronger predictions.

Trend detection is the focus of (CHAE et al., 2012). Their approach provides scalable and interactive social media data analysis including the exploration and examination of abnormal topics and events within various social media data sources, such as Twitter, Flickr and YouTube. Their work is dependent on an analyst to extract major topics from a set of selected messages and rank them probabilistically to then apply seasonal trend decomposition together with traditional control chart methods to find unusual peaks and outliers within topic time series.

In addition to processing the collected data, most of these works rely on showing static plots to display the values of the metrics they are concerned about. As for trend detection, the tools must process information in real-time. For example, TwitterMonitor (MATHIOUDAKIS; KOUDAS, 2010) produces a webpage reporting recent trends in real time and provides an interface for users to rank trends according their own criteria. Since this research is more closely related to the goals our work is attempting to reach, we describe it in more detail below.

2.1.1 TwitterMonitor

TwitterMonitor (MATHIOUDAKIS; KOUDAS, 2010) is a system that identifies trends on Twitter in real time and attempts to provide meaningful analytics that synthesize an accurate description of each topic, and relies on user interaction to order trends by different criteria and to submit descriptions for each trend. The solution is separated in two steps:

- *bursty* keywords detection and
- grouping of keywords into trends based on co-occurrences.

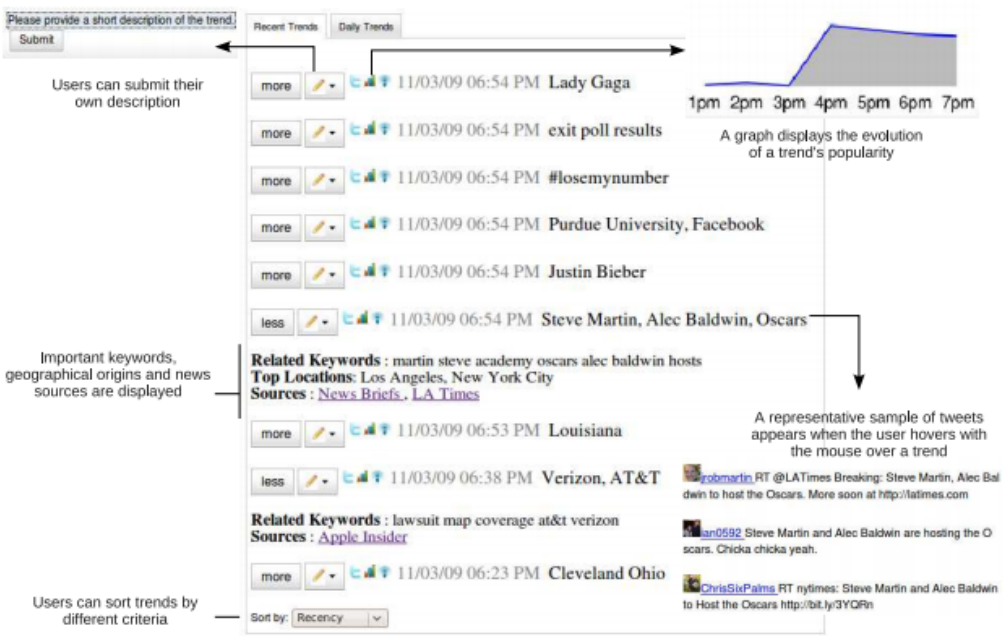


Figure 2.1: TwitterMonitor snapshot. Source: (MATHIOUDAKIS; KOUDAS, 2010)

The authors define a keyword as *bursty* when it is encountered at an unusually high rate in the stream (e.g. the keyword 'NBA' may usually appear in 5 tweets per minute, and suddenly exhibit a rate of 100 tweets/minute), which is frequently linked with an emerging news or event (e.g. an important NBA match taking place). The system considers these bursts as indications that a new topic has emerged and seeks to explore it further.

The detection of keywords is performed by the `QueueBurst` algorithm, which entails:

(i) One-pass. Stream data need only be read once to detect when a keyword is bursty.

(ii) Real-time. Identification of bursty keywords is performed as new data arrives. No optimization over older data is involved.

(iii) Adjustable against 'spurious' bursts. In some cases, a keyword may appear in many tweets over a short period of time simply by coincidence. The algorithm is tuned to avoid reporting such instances as real bursts.

(iv) Adjustable against spam. Spam user groups repetitively generate large numbers of similar tweets. The algorithm is tuned to ignore such behavior.

(v) Theoretically sound. `QueueBurst` is based on queuing theory results.

After a set of keywords is produced, a second algorithm (`GroupBurst`) assesses their co-occurrences in recent tweets, utilizing a few 'minutes' history for each keyword, and when the system deems such activity to be sufficiently representative, it groups them together. Since this process is very expensive in real-time, the algorithm pursues a greedy strategy that produces small groups in a limited number of steps.

Finally, TwitterMonitor analyzes the resulting trends by identifying more keywords associated with a trend, i.e. keywords that do not exhibit bursty behavior themselves but are often encountered in the same tweets as the bursty ones. For this step, context extraction algorithms using dimensionality reduction techniques such as Principal Components Analysis and Single Valued Decomposition, (DEERWESTER et al., 1990) are employed over the recent history of the trend, and the most correlated keywords are reported. Grapevine's entity extractor (ANGEL et al., 2009) is also used to identify frequently mentioned entities in trends, frequently cited sources and frequent geographical origins, and adds them to the trend description.

As a result from the data processing, a chart is produced for each trend, depicting the evolution of its popularity over time. The chart is updated as long as the trend remains popular.

2.2 Visualization of Diffusion Patterns

Besides the many tools that provide graphical ways for monitoring social media activity, there are recent works that propose the visualization of information diffusion patterns (KWAK et al., 2010; YANG; COUNTS, 2010a; CHAE et al., 2012; SUN et al., 2014; LU et al., 2014; WU et al., 2014; ZHAO et al., 2014b; ZHAO et al., 2014a). Among them, techniques described by Sun et al. (SUN et al., 2014), Lu et al. (LU et al., 2014), Wu et al. (WU et al., 2014) and Zhao et al. (ZHAO et al., 2014b; ZHAO et al., 2014a) are the closest ones to a full solution, and we will restrain ourselves to briefly describe them.

Sun et al. (SUN et al., 2014) and Wu et al. (WU et al., 2014) aim at analyzing topics competition - defined as the various topics that exist in an environment that offers limited public attention and, therefore, these topics compete for public attention and media coverage -, in social media (most notably, Twitter) and identify (i) who exerts the greatest influence on a highly cooperative topic - meaning *topic i* and *topic j* are similar and can be divided into cooperation recruitment (attract followers from other topics by the topic leader groups) and cooperation distraction (distract followers from *topic i* by the topic leader groups advocating *topic j* and *j*'s cooperative topics on the followers of *topic i*) - defined that used to be a competitive topic, (ii) what are the similarities and differences in the roles of groups of topic close followers and (iii) how often they divert attention to other topics. Their tool summarizes dynamic topic competition and compares topic leaders to topics by utilizing the Theme River technique described by Havre et al. (HAVRE; HETZLER; NOWELL, 2000). However, it is focused on the patterns of cooperation versus competition of different themes and does not provide support for the analysis of individual profiles.

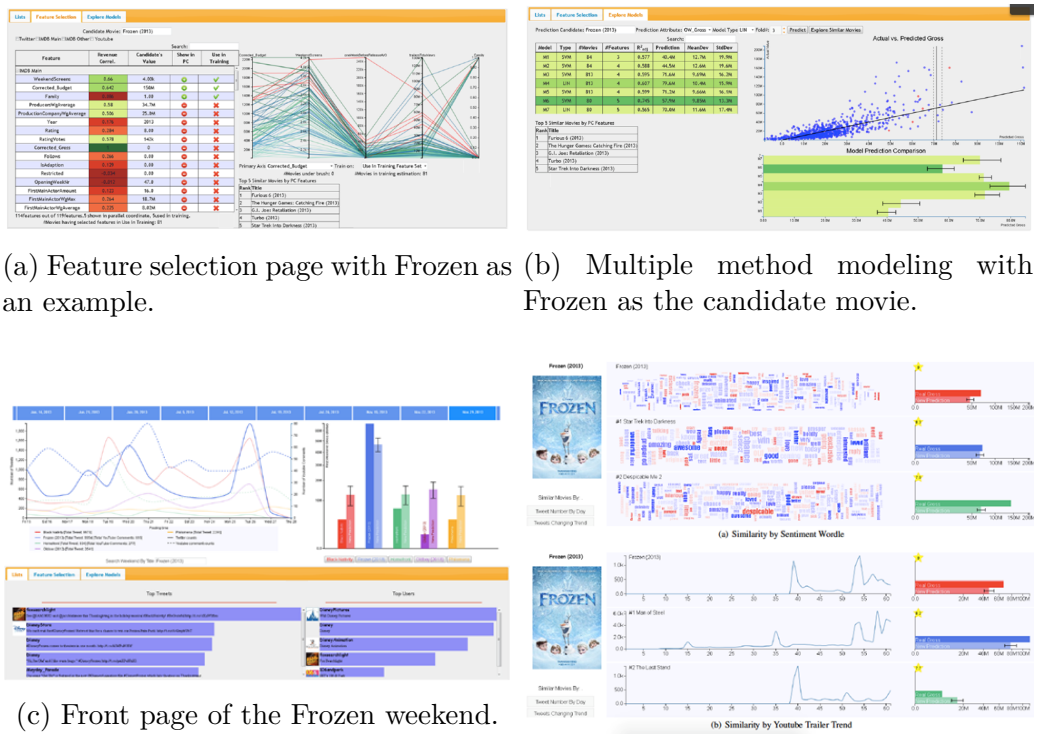
Lu et al. (LU et al., 2014) describe a framework for predictive models using social media (IMDB.com, Twitter and Youtube) in an attempt to create a tool that enables non-domain experts to be able to achieve similar analysis results as experts in a given area. The tool combines line, bar, bubble and candle charts, parallel coordinates and a tag cloud with sentiment analysis as means for the user to explore the information available in the mentioned media and choose from calculated metrics to predict the popularity of movies in their opening weekends. Even though they provide a myriad of visualization techniques, their solution does not enable the user to identify patterns for the increase or decline of popularity of a given subject.

The work of Zhao et al. (ZHAO et al., 2014b) is a comprehensive tool for analysis of emotion and ultimately mood of a given person in social media (most notably Twitter) over time. They provide a multi-dimensional emotion analysis tool with the ability of extracting emotional episodes and infer longer-lasting moods through an enhanced implementation of Havre et al.'s technique (HAVRE; HETZLER; NOWELL, 2000) and rich interaction. However, this analysis is not applicable to press profiles, since these profiles use their microblogs to increase visibility of their online content, instead of commenting on their own day-to-day activities.

Finally, Zhao et al. (ZHAO et al., 2014a) describes a system for the detection, exploration and interpretation of anomalous conversational threads in Twitter. This solution is applicable to the news environment, as we could consider an abnormal increase or decline in popularity of a given profile as an anomaly and apply their algorithms to further explore the available data. The problem lies in the nature of the anomalies, which lacks a clear definition. A thread may be considered abnormal

when it disseminates a message differently from the patterns of other threads in a similar topic. This is not necessarily true, especially when we take channels popularity into account. Common subjects are covered by different profiles in unique ways that would be considered anomalous by their solution.

2.2.1 Integrating Predictive Analytics and Social Media



(a) Feature selection page with Frozen as an example.

(b) Multiple method modeling with Frozen as the candidate movie.

(c) Front page of the Frozen weekend.

(d) Similarity widget view with Frozen.

Figure 2.2: Integrating Predictive Analytics and Social Media: All views of the data matrix visualizations available to the user for analysis. Source: (LU et al., 2014)

As briefly mentioned before, Lu et al. (LU et al., 2014) propose a new framework for movie opening weekend box office gross predictive models utilizing social media. Three different sources are mined for different data: Twitter is used to accumulate comments, time and date, re-tweeting statuses and sentiment; Youtube is used to collect viewers comments on movie trailer videos as well as related videos; and IMDB is queried to define up to seventy two features per movie.

The main objective of the study was to create a tool that would enable non-domain experts to analyze scattered information regarding a given area of expertise and come up with predictions that are comparable to the ones produced by experts. To evaluate the performance of the tool, the researchers used a group of seven graduate students who attempted to predict results for four movies (two with known

results and two with upcoming results). The results were found to be similar to existing expert solutions (BoxOffice.com and BoxOfficeMojo).

The tool was built using known visualization techniques that allow the exploration of the raw data and calculate statistical data in order to compare similar previous results with the titles being analyzed by the user. As shown in figure 2.2a, parallel coordinates are used to display color-coded movie categories and group similar movies with a red hue, also allowing the user to select the features of interest. Figure 2.2b shows the usage of scatter plots to represent Actual versus Predicted gross, and figure 2.2c shows a combination of line charts to display tweets and YouTube comments count over time, bar charts for opening weekend and predicted gross with a candle chart component representing the predicted variation. Finally, figure 2.2d presents the similarities view using Wordle (VIEGAS; WATTENBERG; FEINBERG, 2009) to display the similarities in the sentiments derived from Twitter posts and a line chart to display the comment activity traction over time in YouTube trailer videos.

2.2.2 EvoRiver

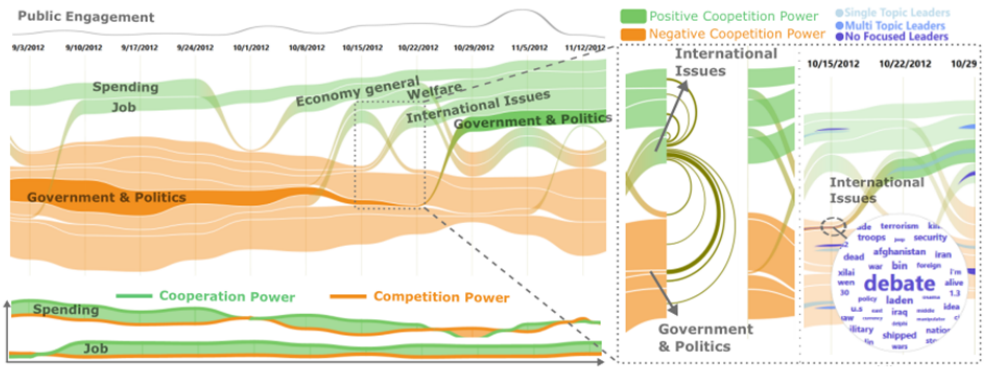


Figure 2.3: EvoRiver. Left image: topic coopetition dynamics during the 2012 U.S. presidential election, showing most of the topics were transiting from competition to cooperation during that time. Bottom image: a chart of spending and job to unfold their coopetition power. Right image: pairwise similarity between international issues and other topics connected by arcs and word cloud of international issues. Source: (SUN et al., 2014)

In this study (SUN et al., 2014), the authors propose a visual analytics platform to allow the users to explore topics of coopetition, defined as the collaboration between business competitors, in the hope of mutually beneficial results. The platform allows data processing and analyzis of Twitter posts as well as interactive visualization. They separate topic leaders in three distinct groups: Single-topic leaders - most active, popular and influential members of single-topic publics -; Multi-topic

leaders - most active, popular and influential members of multi-topic publics, concerned with two to five topics at most; and No-focus leaders - most active, popular and influential members of non-focus publics, concerned about nearly all available topics.

The authors aim to answer six key questions:

1. How does the competition power vary over time?
2. What topics tend to cooperate and compete with one another?
3. Who exerts the greatest influence on a highly cooperative topic that used to be a highly competitive topic?
4. What are the similarities and differences in the roles of groups of topic followers?
5. Do topic followers focus on a few topics? How often do they divert attention to other topics?
6. Can the visual analysis assist in the formation and validation of the hypothesis when patterns emerge?

Based on these questions, the authors derived the the design goals and guidelines. The first goal is to summarize dynamic topic coopetition considering time for temporal patterns and co-evolutionary patterns of topics and topic leaders, facilitate connection of external events and time attribute aids as contextual information. The second goal is to provide a visual metaphor in order to enhance the understanding of topic coopetition. The third goal is to compare topic leaders to topics, by visually relating leaders to each topic. The last two goals are to reduce visual clutter, and highlight and unfold patterns.

In order to fulfill the guidelines, the researchers created the visual representation displayed in figure 2.3, which contains an implementation of the ThemeRiver described by Havre in (HAVRE; HETZLER; NOWELL, 2000). Each stream in the river represents a topic through time, which are colored in green if the activity is classified as cooperation or orange when classified as competition. Streams are grouped in similarly colored large courses and moved when they change from a positive coopetition (cooperation) to negative coopetition (competition) and back. Selecting any stream will display how it relates to the other streams occurring at that moment in time. Opinion leaders are represented with blue hue color over the stream. The opacity of the blue element is determined by the type of leader identified (single-topic, multi-topic or no-focus), and clicking on the element displays a tag-cloud with the words employed by the leader on the Twitter posts used while the publisher was the most popular author discussing the subject.

2.2.3 PEARL

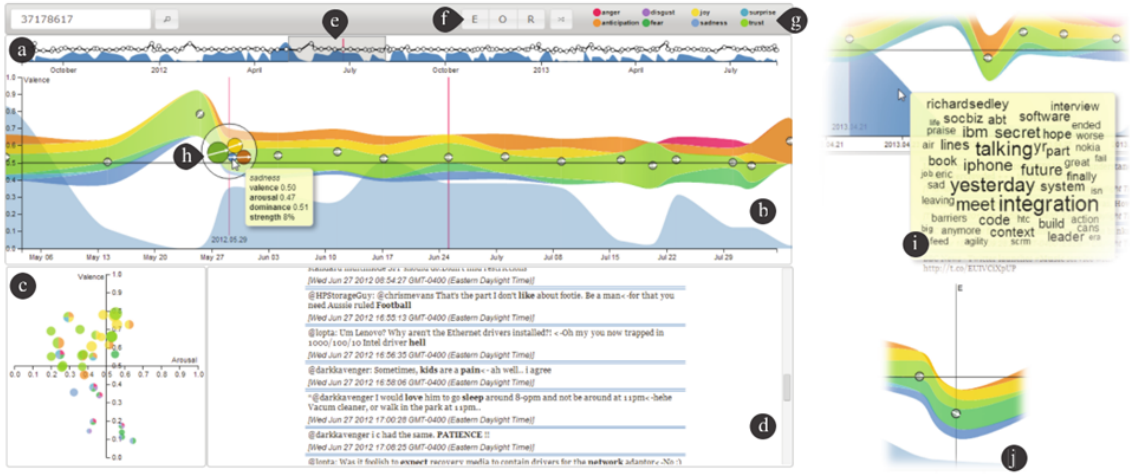


Figure 2.4: PEARL - interactive visualization tool aimed at understanding emotions derived from social media. User interface consists of interactively coordinated views: (a) emotional profile overview, (b) emotional profile detail view, (c) mood word view, and (d) raw tweets view. The overview and detail views are coupled with direct manipulations of (e) a time window. A toolbar on top contains a search box, (f) an action menu, for (j) highlighting important data points, and (g) an interactive legend, for data filtering. It also provides (i) informative tooltips on many of the visualization elements. Source: (ZHAO et al., 2014b)

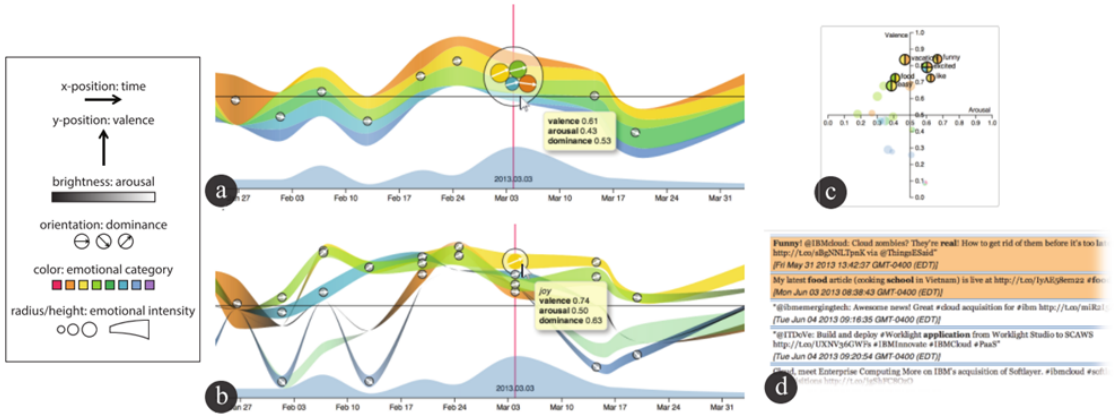


Figure 2.5: Visualization of personal emotions using the metaphors of bands and bubbles: (a) aggregated view, and (b) split view with the green "fear" emotion filtered out. When a user hovers over an emotion bubble of the selected mood, such as "joy" in (b), brushing and linking techniques are applied to (c) the mood word view and (d) the raw tweets view to indicate related visual elements. Source: (ZHAO et al., 2014b)

With this study, Zhao et al. (ZHAO et al., 2014b) aim to answer four key questions:

1. What kind of mood is this person normally in?

2. How often or how easily does he get upset?

3. How quickly can she recover from negative emotions?

4. What usually triggers the person's emotional changes?

The emotion perspective is segregated into three different parts. The first part is the outlook, calculated via the Valency, Arousal and Dominance (VAD) model - Valency, ranging from unpleasant to pleasant; Arousal, ranging from calm to excited; and Dominance, ranging from submissive to dominant - for each word in the Twitter post. The second part is the analysis of volatility of the user's sentiments. The last part is the resilience, measured by the time it takes the analyzed Twitter profile to recover from adversity.

The authors outline their contributions as the creation of a multi-dimensional emotion analysis, able to extract emotional episodes and infer longer-lasting moods, the creation of a visualization rich platform with a rich interaction framework. In order to achieve such goals, the platform filters the tweets based on their content. A tweet is only considered when it's content satisfies a minimal emotion proximity, a semantic proximity and a temporal proximity, meaning the derived sentiments need be similar to the ones observed and pertaining to the same general subject and within a fixed time window. The solution is able to summarize the content analyzed to infer or approximate the cause of the observed user's emotion.

The visualization solutions include an emotion band (shown in figure 2.4-a) that encodes the VAD scale - valency in the Y-axis, arousal in the area brightness, dominance as a white arrow and hues are based in the Plutchik's wheel of emotions (PLUTCHIK, 2001) - and emotion bubbles (figure 2.4-h), where each bubble indicates an emotion component, hue indicates the emotion category, size indicates intensity, brightness represents arousal and arrows represent dominance.

The interaction aspects of the solution are highlighting and filtering to examine emotions, outlook and resilience separately (figure 2.4-f,j) and filter out specific emotions (figure 2.5-b. The second interaction available is the detailing of events on demand, which enable the user to hover in order to bring up a tag cloud of the tweet sentiment and potential emotional triggers (figure 2.4-i), except when hovering a band, which brings up the emotion bubbles (figure 2.4-h). Clicking a bubble brings up the mood word view (figure 2.4-c) and original tweets can also be brought up via the emotion words highlighted (figure 2.4-d). All interaction is coordinated across all views.

There are some key limitations to the solution presented in their work. The study is lexicon-based, but while portable and fast, ignores many linguistic features

when inferring emotions (e.g. negations such as "I am *not* happy"), and the emotion lexicons that PEARL uses were developed based on human-labeled emotional words in a limited context, while different words in different contexts may express completely different emotions. For example, the word “sick” in general evokes a negative emotion, while on social media it may signal completely the opposite as in “Frigging great song. Sick band too.” Given these limitations of a lexicon-based approach, PEARL unavoidably limits its emotion analysis accuracy. Scalability is also an issue since it is not realistic to assume it is possible to process a person’s all social media at once if they have a very large set of linguistic footprints (e.g., hundreds of thousands of tweets or other types of text). To handle such cases, PEARL needs to be improved to support incremental analysis of text input to handle large, changing data set with reasonable response time.

2.2.4 FluxFlow

Figure 2.6: FluxFlow interface with four interactively coordinated components, including a cluster view, a multidimensional scaling view - used to provide a 2D overview of thread distributions in the feature space, where thread similarities are revealed by the 2D distances between them -, a threads view, and a detail information panel. Source: (ZHAO et al., 2014a)

FluxFlow (ZHAO et al., 2014a) aims to enable detecting, exploring and interpreting anomalous conversational threads in Twitter through data processing, a storage

module and a visualization module (figure 2.6). According to the authors definition, a thread may be considered to be abnormal when it disseminates a message differently from the patterns of other threads in a similar topic. The work applies the one-class conditional random fields (OCCRF) model to detect the anomalies, since the data is of the one-class nature (CHANDOLA; BANERJEE; KUMAR, 2009), i.e., little knowledge about true anomalies, and highly time-dependent structures (the user re-tweeting behaviors).

2.2.5 Comparative table

We select six of the aforementioned works that are most closely related to our research, tabulate the objectives described by the authors and the techniques utilized by them to reach them.

Table 2.1: Related work objective and technique comparative table

Objective	Twitter Monitor (2010)	Integrating Predictive Analytics & Social Media (2014)	Pearl (2014)	EvoRiver (2014)	Opinion Flow (2014)	FluxFlow (2014)
Topic evolution	Textual	Line chart		Stream graph, tag cloud, line chart	Stream graph, tag cloud	Bubble chart
Prediction		Bar chart, candle chart			Bar chart	
Popularity		Bubble chart				Connected graphs
Sentiment		Wordle	Stream graph, bubble chart		Stream graph	
Search for influential profiles					Connected graphs	
Topic leaders				Stream graph, tag cloud, line chart	Stream graph	
Anomaly detection						Bubble chart

We grouped the objectives in seven major categories:

- Topic evolution: the manner each topic in the set changes over time
- Prediction: attempt to allow the user to make intelligent predictions based on the available data and the patterns previously found
- Popularity: how the profile's popularity evolves through time
- Sentiment: the analysis of the sentiments involved in a set of messages and how such sentiments can be used to find interesting patterns
- Search for influential profiles: identifying separate profiles that help increase the traction of the original profile's messages

- Topic leaders: how each topic struggles to obtain more attention from the readers and how the topics reach the main focus of its readers
- Anomaly detection: identifying how regular conversation patterns occur and which topics behave differently than the expected behavior

Through this analysis, we identify the main interest of the research is to detect topic evolution and sentiment analysis using mostly known techniques like stream graphs, bar charts and bubble charts. The repeated usage of these techniques indicated a familiarization of the users when representing the respective data. We also identify the lack of work regarding geographical representation of the data, which makes it a compelling area to be explored.

With the exception of Integrating Predictive Analytics & Social Media (LU et al., 2014) and Opinion Flow (WU et al., 2014), most research does not attempt to utilize a greater number of complimentary objectives. Instead, most research is focused on a small set of goals, separately.

After studying the state of the art, we defined our objectives including a large number of complimentary goals in order to enable the user - expert or not - to do an in depth research of the data and to understand how each different interaction in the information diffusion process impacts the reach of the profile in social media.

3 PROBLEM DESCRIPTION

We first started our research trying to understand the geographical components involved in information diffusion in social media. We understand there is a particular difference between these two forms of social media: while in Facebook, users (people or organizations) are connected by explicit reciprocity in a "friendship" relation, in Twitter users are connected by a "who follows whom" without the need of reciprocity (KWAK et al., 2010). Despite this difference, in both tools one could measure how far goes some information posted by a user, via cascading shares in Facebook and re-tweets in Twitter. Moreover, the richness of the data sets provided by these tools allows the investigation of the structure of these networks, the patterns of information diffusion along time, the communication dynamics between individuals and among groups, and the role the influential users perform in local or global communities.

We then summarized our interest in four questions:

1. How an information posted by an individual or corporate user spreads over the network? Who consumes and broadcast information from a given source?

2. Which sources are most followed in a given region?

3. How different sources post the same event?

4. Is there a temporal relation between posts from different sources?

The answers to these questions may vary depending on culture, region, the information itself, economics, and providing ways of investigating them is a challenge. We created a solution intended to be generic but restrained ourselves to the use of Twitter data for our first case studies. We proposed several visualization designs to identify temporal, geographic and behavioral patterns of information consumers of different news source spread over social media. The combination of designs revealed hidden patterns in the massive data set of messages and comments worldwide. The results obtained with these designs were the seed for the main problem we address in the present dissertation.

Next section describes our first proposal, facilitating the understanding of our final motivation, which is presented in the last section of this chapter.

3.1 Information Diffusion in Social Media

3.1.1 Visualization Designs

We have developed a combination of visualization designs (Figure 3.1) that complement each other forming a mash of patterns that are able to tell the user when, where and how information is shared among any community of readers from a news source.

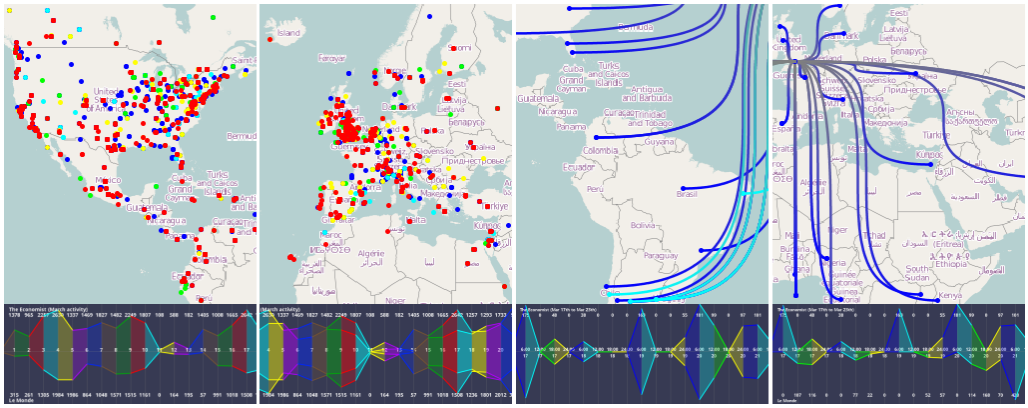


Figure 3.1: Visualizations design. Using color coded geo-located marks (on the left), the designs represent highly invested areas in contrast with uninterested locations, coupled with color coded streams (at the bottom) that compare time spans of high and low activity worldwide. Gradient flows indicate the direction of information in time and space (on the right). Source: author.

3.1.1.1 Data

The data used for the study was taken from Twitter public feeds and was comprised of a text message, the user name of its author, the location of the author (which is a free text, input by the author in their profile), the latitude and longitude where the message was posted if available, the date in which the message was first posted, the original message to which the text message was responding to or re-tweeted from (all from select known news sources worldwide), the original author, the location of the original author, the date and the number of times the original message was re-posted. Figure 3.2 shows an overview of the main processes involved in our approach.

One of the key features of the data for the analysis planned in the visualization

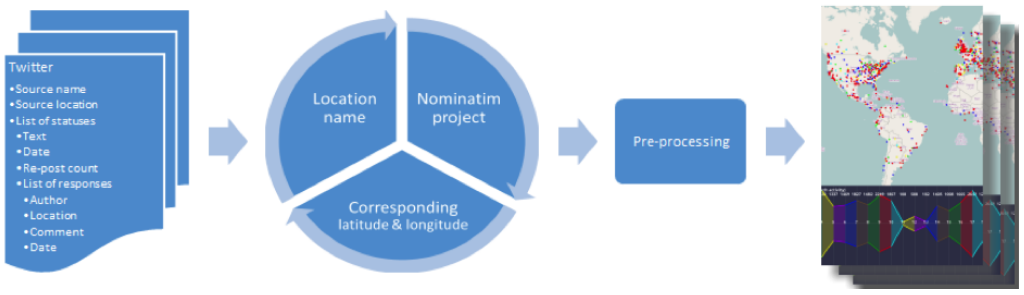


Figure 3.2: Data gathered from several different news sources on Twitter over the internet via Firehose API and filtered to store key aspects to be used in the final analysis. Status location was only available in 5% of the dataset, so for each unique user posting, the location stored in their public profile was also gathered, stored in a static factory and sent in a batch format to OpenStreetMap Nominatin API via HTTP post. The response was also stored in a static factory for use in pre-processing. During processing, number of re-tweets are accumulated to each status with the comment and post date and finally passed on the the GUI to be rendered as alternative visualization designs. Source: author.

tool was the geographic location of each post and, though it was possible to include the exact latitude and longitude of each message on the moment of posting, only a little over 5% of the messages captured during the study were accompanied by a geographic location. In order to mitigate the losses, OpenStreetMap (HAKLAY; WEBER, 2008) was used during data parsing to translate the free text input by authors as the main location of their profile into usable geographic points (latitude and longitude).

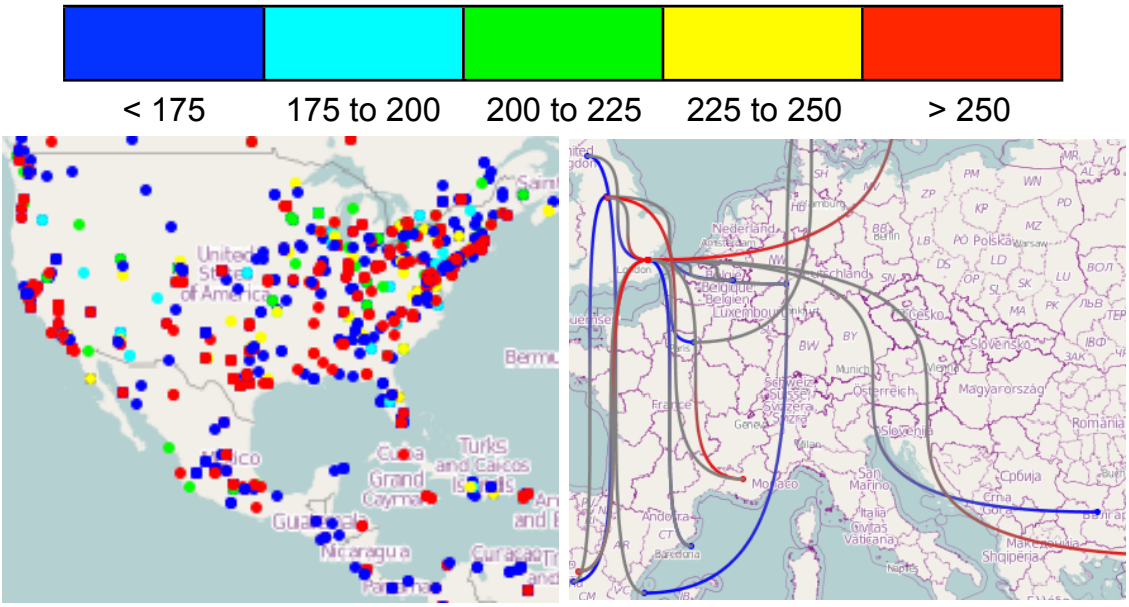
Prior to rendering the visualization, several calculations to the gathered data needed to be applied in order to define the metrics the final tool revolved around. One of the most important and perhaps the one contribution that greatly differentiated this work than any other found in literature was the search for community behavioral patterns based on genetic algorithm applications. With said algorithms, it was possible to determine, even before any rendering, the best candidates for diffusion patterns found in the time series. By adapting the algorithm to interpret social media messages, we were able to determine such proposed areas of interest where repeating time series patterns could be found and categorized their quality by the size of pattern and how many times it repeated itself both in one time series and in multiple sources.

3.1.1.2 Technique description

The visualization tool consisted of two separate designs that complemented each other by giving two separate interpretations to the same data at any given time

during the animation process, one from a geographical standpoint and the other exploring activity over time.

The first design (Figure 3.3) was comprised of a map with the Mercator projection as a background on top of which each source and reader was rendered. Sources and readers were rendered with a small mark on the location described in their Twitter profile with a color code representing the level of activity on that location, based on the number of re-posting and commenting of original statuses. Curves connected the network of re-tweets displays which posts were published first, making it obvious how long it took to any location to take notice of the news posted.



(a) re-tweeting activity (b) temporal flow

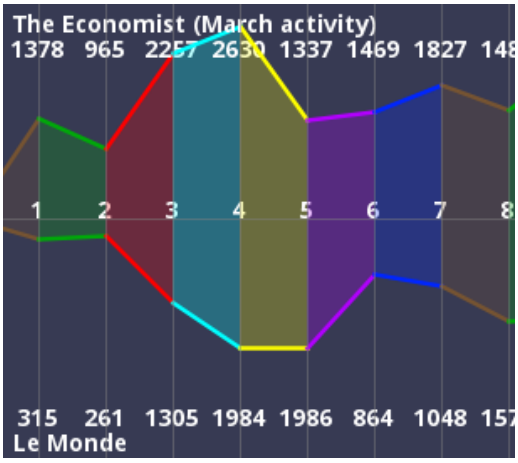
Figure 3.3: Map designs. Each separate location was marked with a color coded dot, according to the number of times the original message was re-tweeted. Different zooming levels revealed the information in more detail and filtered out information of lesser interest. Temporal flow connected locations to each other according to date and time of the post with directed lines. The curved lines left as gray from the earlier post location and arrived with the corresponding color in the later post location. Source: author.

Interaction with the visualization could be executed by mouse and keyboard. Hovering any location would display the original tweets being commented on as well as the absolute number of re-tweets. Clicking a location selected the original messages, faded the unrelated posts and displayed the aforementioned network for analysis. Typing any letter filtered out any message that did not contain the input in their original text, which made it easy to search for subjects of interest for in-depth analysis.

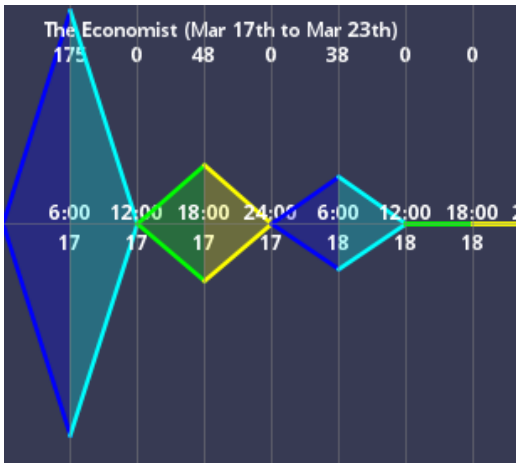
The second module (Figure 3.4) was comprised of a stream graph representing

the number of re-posted messages of each visualized source over time, in the same hue as their mirrors in the first module. The user had the option of overlaying one stream graph for each source or use the same graph for two different sources. The level of detail can also be parameterized for the optimum time series analysis. With this implementation, it was extremely easy to identify the reader community reaction in both gross quantity and intensity to each message. This feature made it possible to compare news sources with different reader densities to be brought to a leveled playing field and interpret how much of an impact they are in their respective communities, since even though gross numbers may vary, the patterns in inclination of growth and reduction in responses could be easily found. To that point, the motifs discovered in the pre-processing algorithms were marked with a brighter saturation in the area they represent, so that immediate attention was called to that part of the data, and interpreted as a pattern found in the reading community of either source.

Hovering any source in the second module had the same behavior as discussed in the first module, in that hovering any stream called attention to its referred source.



(a) Activity by day of the week (Sunday is colored red and Saturdays is colored green)



(b) activity by 6-hour shift (12AM to 6AM colored in dark blue and 6PM to 12AM in yellow)

Figure 3.4: Stream graph design. Activity separated by time or date window showing the pattern of increase/decrease from each time span to the next. (a) Activity of March, 2013 of two different sources (The Economist on top, Le Monde on the bottom), by day of the week - each day with itś own color. (b) Activity from a single source in the week of March 17th 2013, separated by 6-hour shifts. Chosen spans are marked with a light gray vertical line along which the absolute number of re-tweeting from each source is displayed, as well as the limit of the chosen span. Source: author.

3.1.2 Implementation

The entire project was implemented in Java using AWT library to render the visualization, Twitter4J API to gather all data and OpenStreetMap Nominating to translate location names into latitude & longitude points. In order to improve performance, locations were gathered and stored a static factory design pattern, which prevents the need to process previously used locations. This technique improved not only performance, but also memory usage. Consultation of OpenStreetMap Nominatim project were made via batch at scheduled times, so that the number of new locations could be maximized, minimizing the number of HTTP Posts. A common problem in the field of geographical visualization is the translation from latitude & longitude pairs to 2D canvas points, due to the nature of the Mercator projection. We solve this issue by implementing a translation algorithm similar to WGS 84(TURNER; ELGOHARY, 2013). This algorithm, however, has an intrinsic problem of accumulated error of about 0.5%, which we accounted for in this implementation.

3.1.3 Case Studies

Several case studies were created in order to test the effectiveness of the visualization tool regarding two key aspects: geographical impact of a news source; and community reaction over time. The hypotheses created to support the analysis in these aspects were:

1. Global news sources would have a broader audience worldwide and local news sources would have a more localized audience;

2. Global news sources created more reaction while posting news regarding global issues while local news sources had a more expressive reaction when posting news that important to the local community and/or to emigrants from the local community;

3. Globally known news source reading communities would have a more immediate response then local news sources;

4. Local news source communities would have peaks and valleys on their activities while global news source communities would have a more steady behavior.

These hypotheses served as indicators to which type of news can gather more followers as well as which hours of posting result in better response from the reading community.

In order to support the findings, two different types of news were used: the first one is the election of Pope Francis, used as a control to analyze the reaction to news of global importance, covered both by global news sources and by localized sources; and the weeks preceding and following the week of the election, which had little news of such global impact, meaning each source was observed during normal activity.

3.1.3.1 Case Study 1: Global Impact

Figure 3.5 illustrates the difference in community activity regarding news posted from two major European sources over a 1 month spread. Immediately, it was possible to identify a few trivial patterns such as a larger activity concentrated near the news source and on locations closer to the ocean worldwide. It is also noticeable that The Economist (top source) had a more vocal community, in contrast with Le Monde (bottom source) with a moderate level of activity. The figure reveals a massive interest from users located in the United States in the foreign publications. In fact, pre-processing revealed that 65% of the community activity measured via number of retweets came from North America alone, dispersed over an average of 335 locations per day in a given week, which makes this an interesting case for further analysis.

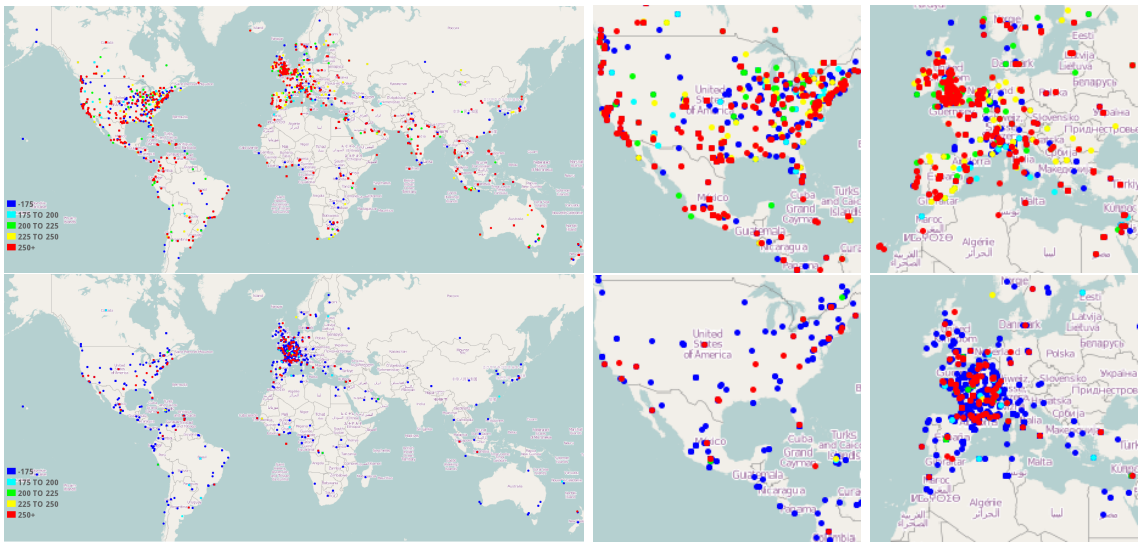


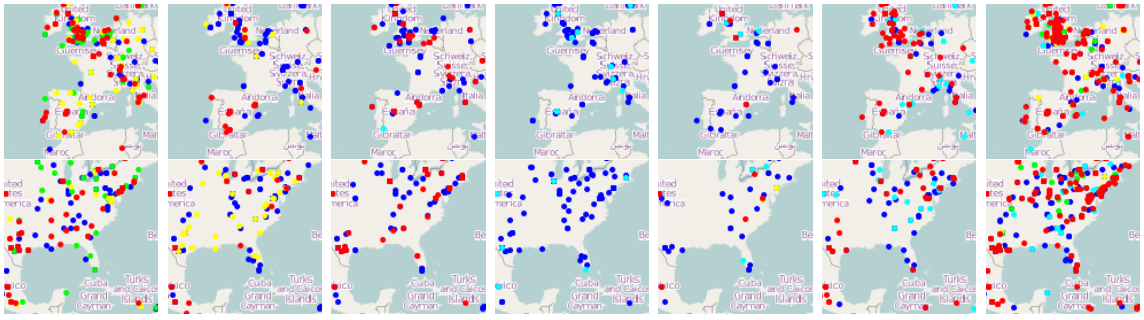
Figure 3.5: Global Impact: two major European publications visualized according to number of re-posted items over a 1 month period. Top left: the world view for the Economist. Top right: two detailed zooms over North America and Europe. We observed a larger spread with a bigger community response. Bottom row: results for Le Monde, activity mainly in western Europe. Source: author.

3.1.3.2 Case Study 2: Days of the Week Impact

Figure 3.6 reveals an interesting pattern developing from the analysis of days in a week separately. The 14 images represent the community activity related to news posted on each day of the week (Europe on top, United States on the bottom), regardless of the time the reader re-posted or commented on the tweet, starting on Sunday and going all the way to Saturday. With this technique, it becomes clear that posting a status on Friday, Saturday or Sunday tends to have a better response from the community, in general. This pattern was found repeated on most sources during a span of two and a half months of gathered activity. North America region was of special interest, since it represents over 65% of all activity from all sources gathered for the construction of this tool. We can clearly see that the pattern found in Europe repeats itself not only in the number of re-tweets, but also in the number of locations showing activity. The lowest point of interest shown by the reading community tends to be on Thursdays, where the activity decreases about 80% from the average of 335 locations in North America. The decrease of re-tweets is also very significant, almost reaching 40% from the average 4500 re-tweets per day. Wednesdays have a very similar behavior, though not as drastic from the number of locations standpoint.

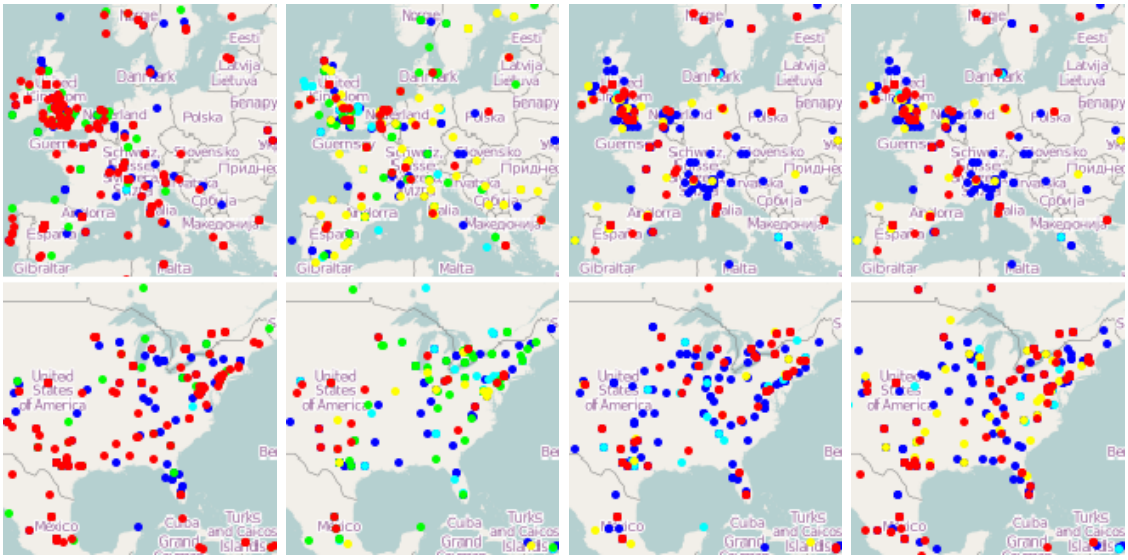
3.1.3.3 Case Study 3: Daily Shifts

After analyzing each day of the week and noticing a pattern in the distribution of activity in those days, it became apparent that interesting patterns could arise from the analysis of 6-hour shifts separations as well. Figure 3.7 displays this case study, and reveals an interesting pattern: news posted during the night, even during the period after midnight and before 6AM, seem to be the most attractive, since these



(a) Sun (b) Mon (c) Tue (d) Wed (e) Thu (f) Fri (g) Sat

Figure 3.6: The Economist over a week, zoomed into Europe and the United States. (a) and (g) are responsible for over 40% of all activity and almost 60% of all locations Source: author.



(a) 12AM-6AM (b) 6AM-12PM (c) 12PM-6PM (d) 6PM-12AM

Figure 3.7: The Economist over daily shifts of 6 hours, zoomed into Europe and the United States. (a), (b) and (c) are responsible for. Source: author.

are the messages that generate the highest levels of activities, regardless of when the re-tweeting or commenting takes place. While using the same parameters to analyze the behavior in North America, the number of responses denotes the same activity found in Europe, although the number of locations show small variation, with the exception of the 6AM to 12PM period, which shows a considerable decrease in the number of locations consuming news items from the source. Another interesting aspect is that the period 12AM-6AM shows a smaller contingent of locations, despite maintaining high levels of activity, which means that even though fewer people are reading the news from these periods, the ones who do read are more vocal and are responsible for the vast majority of the re-tweeting. These patterns can be explained by the difference in time zone from the The Economist and Le Monde locations and the other locations worldwide.

3.1.3.4 Case Study 4: Network of Re-Tweets for Individual Topics

After analyzing re-tweeting activity over a month in its entirety, some interesting facts were detected that threw a few numbers off the curve. Figure 3.8-left is the representation of all tweeting activity regarding the election of Pope Francis, on March 13th 2013. By analyzing the week during the election of the new Pope, we detected a valley on the activity for one of the most consistent source found (The Economist). After analyzing other sources, it was apparent that even though this magazine reported on the event, it was more largely covered and followed via

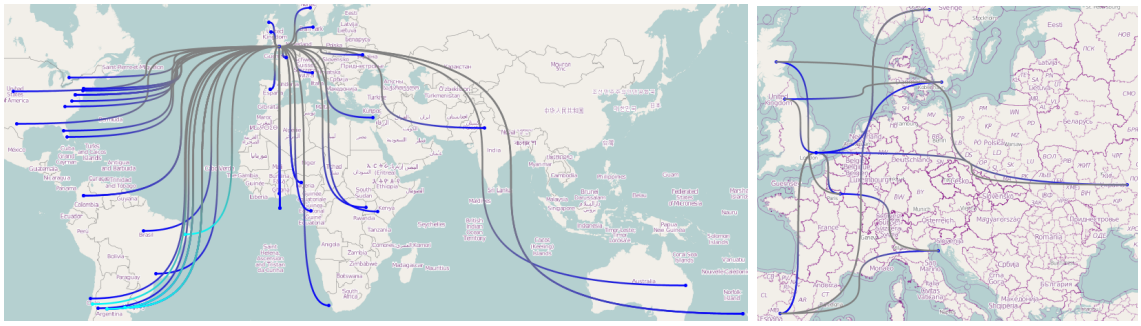


Figure 3.8: All messages regarding the election of Pope Francis over the world (left) and in Europe (right). 443 re-tweets total (139 in Europe) spread over 59 sources (14 in Europe). Time related flow mapped via directed curves on the bottom map, symbolized by a color gradient starting as gray, near the source, and finishing with the appropriate color code when approaching the activity location. Source: author.

other more popular sources, such as The New York Times, which demonstrated spikes in the normal activity. Even more interesting than this, South America demonstrated a spike in interest and was the only sub-continent to break the 175 re-tweet barrier in Argentina, Brazil and Chile. In the enlarged corner, we can see South America with time flow mapped from each re-tweet to the next, so we can understand which location first reacted to the original message. Interestingly, even though Pope Francis is Argentinian, the original Pope related news from The Economist was first re-tweeted in Brazil, then Buenos Aires and Mar del Plata, then Chile. This phenomenon is explained by the fact that Brazil is the country with the largest catholic population in the World and, therefore, should have proportionally more readers following the story. In Europe, after the original messages are posted, the regular flow occurs by reaching northern England and Wales, then Sweden, and on the southern part of the map, to Spain then Italy and Ukraine. It is clear that even in London, this piece of news was of much interest via this source and, instead, looked for news sources closer to the event (such as @news_va_en) or a wider source such as the New York Times. Figure 3.8, on the right, is a similar case, with the subject of the messages being the BRICs (Brazil, Russia, India and China). This analysis is a great contrast to the previous one since the countries most directly related to the news have little if any activity. India, China and Brazil have one location in their area each and Russia did not show any activity. China's activity (or lack thereof) is easily explained by the fact that Twitter is now allowed in the country, with the exception of Hong Kong. Russia as whole is not a very active follower of The Economist, but Brazil and India have consistently shown a moderate to large activity, though in this case, very little interest was shown. In Europe, the response pattern was consistent with other general information already detected, although the number of re-posting was significantly higher.

3.1.3.5 Case Study 5: Stream Graph Design Analysis

The second design proves an excellent source of temporal patterns in information diffusion. Figure 3.9 shows that even though the absolute numbers in The Economist are higher, Le Monde displays similar patterns of growth and diminishment over the week, with the exception of Wednesdays, which consistently display growth while the Economist consistently displays a small diminishment. Additionally, it becomes very clear that during the week of the Pope election, both sources had very little activity going on. When digging deeper into the source by using a more precise granularity, it is possible to see that activity during the night greatly dominates activity during daytime, which points to the fact that even though the followed source is a recognized source of financial news, most of the interesting news are posted during the hours the market is closed and are, therefore, not representative during working hours.

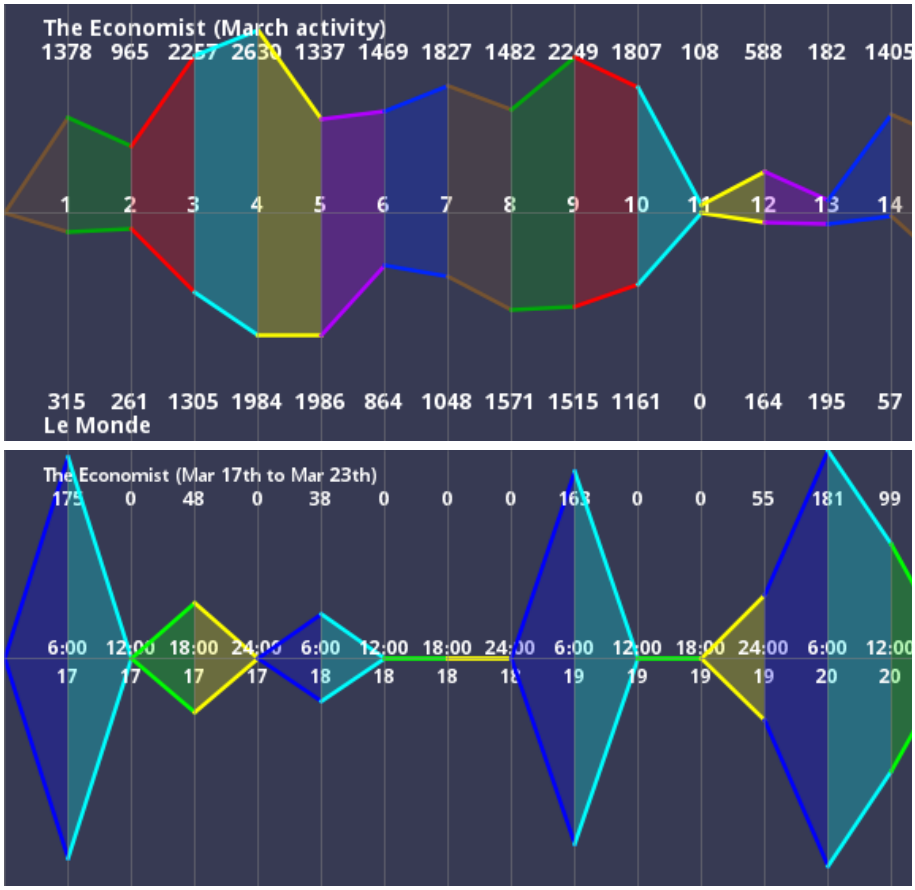


Figure 3.9: The Economist compared with a second source and isolated in a single graph. (a) The Economist (top) vs. Le Monde (bottom) visualized as Stream Graph. Sundays are drawn in red and the week follows. It shows a larger activity than its peer, but a similar trend. (b) Smaller granularity, the week of Mar 17th to 23rd separated in 6-hour shifts. Source: author.

3.1.4 Evaluation study

An informal evaluation session was performed with eleven subjects of varied ages, education and experience. We describe the sample of subjects in 2 tables, one for the general characteristics (table 3.1) and one for the subjects' experience regarding visualization techniques (table 3.2).

The sample of eleven subjects was composed by 9 males and 2 females, 7 under 30 years of age, 2 over 30 years of age and one did not wish to disclose his/her age. Among them, 1 was an undergraduate student, 8 were M.Sc. students and 2 have already graduated, all of which had a computer science background. Two subjects had no experience, 7 had little experience and 2 were experienced in general visualization techniques. Regarding visualization techniques targeted towards social media, 4 subjects had no experience, 4 had little experience and 3 were experienced.

They were asked to use the tool with the data set comprising re-tweets from The Economist and Le Monde during March 2013, i.e., the same data set we employed in the case studies. During a one hour session they evaluated the map design and the stream graph, trying to answer the following questions:

1. Is there any difference between Le Monde and The Economist in the way both report the pope Francis election?

2. What location had the larger number of re-tweets in this case?

3. Do you recognize any peculiarity in the flow lines in Europe when filtering the data set using "BRICS"?

After the experiment they answered a Likert-scale questionnaire, with 12 questions assessing the key visualization designs' attributes. All subjects agreed that the color scale is adequate in the representation and that the designs help to easily identify regions of high and low activity. Most users agreed that the stream graph makes it easy to compare multiple sources and to spot specific behaviors related to the days of the week. Most users agreed that it is not intuitive to connect the map and stream graph designs in a complementary way and that the legends in the stream graph design were not informative.

Table 3.1: General characteristics of the evaluation session subjects

Sex		Age			Education	
Female	Male	Under 30	Over 30	Undisclosed	Undergraduate	Graduate
2	9	7	2	2	1	10

Table 3.2: Visualization experience characteristics of the evaluation session subjects

General Visualization Techniques Experience			Social Media Visualization Techniques Experience		
None	Little	Experienced	None	Little	Experienced
2	7	2	4	5	2

3.1.5 Improved analysis experience

The techniques explored in the designs proved to be of great value on the analysis of information diffusion. Among other qualities, four important types of patterns can be intuitively detected with this combination of designs. Geographic patterns of information diffusion are easily found by looking at the disposition of color coded marks on the map. Higher activity locales immediately pop to the front, making it easy to understand where the interest is concentrated. Usual geographic visualizations either use the size of the mark to convey the activity level, which increases occlusion of high traffic areas, or color coding an entire location (i.e. United States or Sweden), which inadvertently calls more attention to locales with a larger area then the ones with higher activity, specially in regions such as Europe of the Middle-East, which are small but filled with different high activity locations. The proposed map also improves the ability to compare locations from one region to others in completely different regions, as opposed to conical, pseudo-conical and azimuthal projections, which only show a portion of the globe at a time.

The proposed activity flow enables the user to understand how information is passed from one reader to the next and, in many cases, back to the source. This type of analysis can be used to understand which location has the higher interest in determined subjects as well as which locations are used as hubs for the diffusion of information. Locations that constantly appear between the source and determined readers can be classified as hubs in the diffusion since end readers only comment or share information after it is declared as item of interested of these middle locations. Flow can also be used to understand if a location is interested in the source or in the subject of a subset of the news posted by the source.

The use of stream graph coupled with discrete sampling in determined windows enable users to understand time-related patterns in the information diffusion process. By analyzing peaks and valleys in the stream, it becomes easy to find which time spans repeat their behavior consistently and can, therefore, be used as best times for posting news the source is highly invested in, to maximize the response from the community. Additionally, all cyclical patterns can be identified intuitively and compared with multiple sources to categorize the activity between community related or source related.

Timely responses can be identified and content can be tailored to specific subsets of readers by analyzing the peaks and valleys in the streams. These key roles are very important to know since they play an important part in the diffusion of information via social medias. Despite the robustness of established news source, the internet has an ephemeral quality that awards highly invested readers that frequently comment quickly and objectively on different subjects and are held in high esteem by the community. This information is also key in the tailoring of information delivery.

3.1.6 Limitations

With any visualization design, occlusion is an important issue that may impair the quality of the final results. By the nature of the data involved in this project, there is a discretization of the used coordinates, which is increased by the usage of free text to inform the location from which activity comes and, therefore, the quantity of occlusion. Having a static format such as the Mercator projection makes for a limited space for creating appropriate analogies, since highly important locations are not necessarily far apart from each other and can easily become cluttered with information, such as Europe and the United States, in the case of the data set used in this work. Both of these problems may be averted by filtering data with more flexibility. Stream graphs are limited by our ability to differentiate between too many different lines and widths. Our experiments also showed us that comparing more than five sources in a single stream graph becomes impractical.

3.2 Specialists Inputs and Decision Making Process

After we concluded our research and analysis in the previous work, we were interested in investigating other aspects of information diffusion, so we contacted experts to understand the problems they faced in their day-to-day work regarding information in social media.

We met two journalists from different major newspapers (Folha de São Paulo and Zero Hora), and we presented our work in separate occasions. After discussing the problems we had researched in the past and how each company uses Twitter to increase traction for their publications we began an interview in which we asked multiple questions regarding the utilization of social media (focused on, but not limited to Twitter). The questions were categorized as

1. Importance of Social Media in popularization of stories

2. Analysis of results from stories promotion via Twitter and

3. Evolution of Social Media usage in popularization of stories inside the company.

After the interview, each specialist then went back to their offices and had a week to come up with a set of questions they would like to have answered about the readers and their behavioral patterns. After a week, they sent us a set of questions. We analyzed them and summarized the results into a single cohesive list, which became the foundation for our final research. The list was then shared with the experts for approval. The final list and expert comments follows (key questions marked in bold typeface):

1. How many posts are shared by verified profiles? *After review, found not to be interesting enough by the experts, so we eliminated the question from the research*
2. **Why do users stop following a given profile?**
3. **Does the time of the post affect the traction to the story?**
4. How does a profile relate to its followers?
   (a) Which words are used the most?
   (b) **What time are the stories tweeted?**
   (c) **What time do the re-tweets occur?**
   (d) **Which sentiment is associated with the profile?**
5. Do trending topics occur before or after the original posts from the source covering the subject?
6. **Is there a co-relation between the subjects of the posts and the popularity of the source profile?**

From the final list, we determined that the visualization techniques utilized by our solution needed to be simple enough that the end user would be able to utilize them to explore a known data set of messages with minimal support. We chose the combination of bar charts, to plot the changes of popularity of selected sources, bubble charts to plot the popularity of each single post versus the time of original post, tag clouds to represent the words utilized by the sources and by the readers, as well as the sentiments involved of the messages in which the words were present, a panel containing each post in its entirety as well as information about the author and the sentiment of the post and a configuration panel to select from the available sources, analysis features and visualizations.

After we created the first working version of the tool, we shared it with the experts and conducted a second interview. During this process, we first presented the solution and shared details of the available data set, techniques and analytics, then we let them play with the tool, answering questions as they were posed. We refrained ourselves from guiding the experience, since we wanted the final implementation to be as intuitive as possible for the end-user. Each journalist had no limitation of time, but all sessions were held at the same workstation in order to prevent machine configuration bias while evaluating performance.

Next chapter describes our solution in detail.

4 VISUALIZING INFORMATION DIFFUSION IN SOCIAL MEDIA AND HOW PRESS INTERACTS WITH SOCIAL NETWORKS

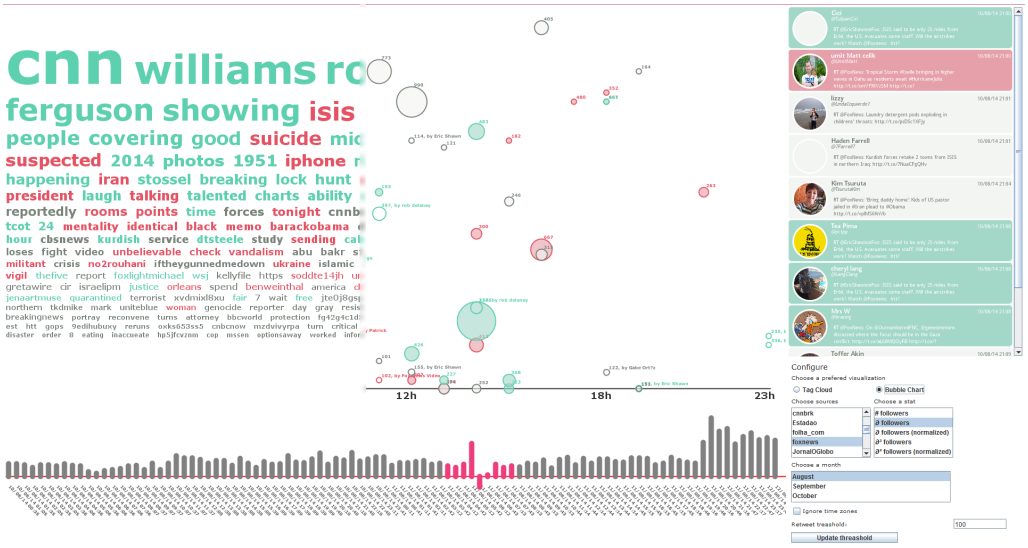


Figure 4.1: Overview of our interactive solution that combines different visualization techniques: tag cloud, bubble chart, bar chart and message board. On the bottom right corner, the configuration panel. Source: (BOETTCHER; COMBA; FREITAS, 2015)

We proposed the use of a combination of bar chart, bubble chart, tag cloud and message board (figure 4.1) coupled with responsive interaction among the combined visualization techniques to bridge the gap left by other studies in the same area. The usage of a combination of different visualization techniques allows co-relating the activities of a profile and of its reading community with the resulting popularity. The solution provides for multiple source comparison, so that not only the user is able to understand their own community but also the success and pitfalls faced by the competition in the same medium. Through interaction, it is possible to reveal hidden patterns in the data set of messages and comments worldwide, enabling the users to have unique insight into their communities' behaviors and preferences.

We combined analytical and statistical information from original posts by news profiles and readers alike, channel popularity information and trending topics. We included configuration options that allow the user to filter out any activity that falls outside of the subject of interest and concentrate on the actions that closely relate to the popularity spikes and valleys found by the user. We took the user's expertise into account when defining which are the abnormal activities in the available dataset, since automatic search for such patterns falls into the nature of true anomalies for which there is no clear definition.

In order to gain insight into the daily work of the target users for this system and how it could help them in obtaining better results from their efforts in social media, we conducted interviews with journalists from different major newspapers (refer to section 3.2). Based on the gathered information, we tailored our solution to answer the key questions posed by the group of journalists:

1. Why do readers stop following a profile?

2. Does the time of the post co-relate with the number of re-tweets?

3. How does the profile relate with its network?

4. Is there correlation between the subjects of the posts and the amount of followers/popularity of the source profile?

After creating a working prototype, we shared our proposed solution and monitored their usage. We mention their findings and analyze their reports regarding the tool in the results chapter.

4.1 Data gathering

Data was gathered from Twitter over the period of 3 months with the use of the Firehose API, by filtering the messages by original author, re-tweeted status author or user mention to match the list of 19 news source Twitter IDs spread worldwide. We built two different tools to allow for a complete gathering of the data.

The first tool is a constant consumer of the Firehose API, which was designed to continuously capture live data, 24 hours a day. This tool would capture JSON objects as displayed in figure 4.2, which was separated into three sections: message data, author data and geographical data. The message data includes information on creation and editing dates for each post, as well as the number of re-tweets and responses to the original message and the list of contributors to the specific post. The author data includes the name, handle, location and number of followers of the posts' author. Finally, the geographical data contains the latitude and longitude

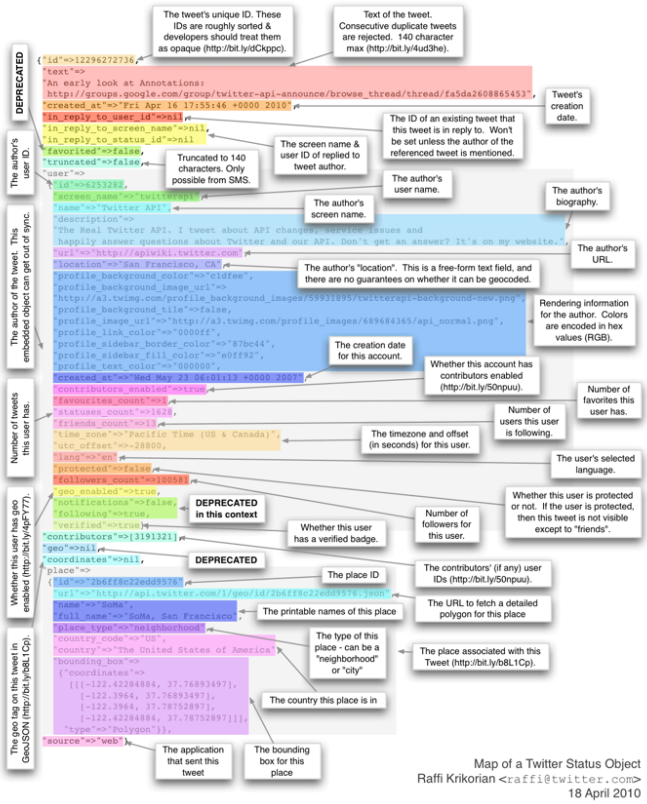


Figure 4.2: Map of a Twitter status object, by Raffi Krikorian. Source: http://www.scribd.com/doc/30146338/map-of-a-tweet

pair as well as place (which is the boundary of the location from which the post was created), when enabled by the author. Such information is rarely allowed by the profile, amounting to only 5% of the tweets we were able to capture. We confirmed this occurrence with information found in several blogs and forums regarding the subject.

The second tool is a daemon process that would look for gaps in the data and utilize the Twitter Search API in order to find the messages missed during the initial gather and complete our data set. The usage of this tool allowed us to account for any errors or glitches in our primary tool, as well as for power outages or connection resets experienced during the data gathering phase.

The total numbers amount to 15 million original tweets from August 2014 to October 2014, averaging 5 million publications per month and over 10.000 posts per source per day. Each status update object is stored raw in a JSON format on plain files (see figure 4.2). Complementing this data set, further data was gathered from each source, containing the number of followers every 30 minutes for the same time period, resulting in an average of over 1000 snapshots per source per month. Finally, a third and separate data set gathered the trending topics from the 15 available locations in Brazil every 30 minutes, resulting in 495 unique trending topics over

the three month period (an average of 165 unique trending topics per month).

4.2 Tag Cloud

Figure 4.3: Tag cloud of words used by news source with color-coded sentiment. Users are able to separate original source material and community re-tweeting and comments and select words of interest. Selecting words will act as a filter in all other views to only display content from messages that include the selected words. Sentiments are colored in green (positive), gray (neutral) and red (negative) and is derived from all messages that contain the word. Source: (BOETTCHER; COMBA; FREITAS, 2015)

The tag cloud shows the most recurrent words and the sentiment of the messages relating to them for the selected profiles during the time period defined by the user.

The visualization is implemented as displayed in figure 4.3, with either the words from selected profiles or with a localized list of trending topics. Like most implementations of this technique, occurrence of the word is mapped to its font size, so that the most commonly used words or tags will be largest as well. Most users will use this visualization to understand which subjects were of interest in the selected window of time, so to augment the importance of each term, the most common words are also listed first in the visualization.

The sentiment attached to each word will be calculated from the messages that contain them. We used the SentiWordNet 3.0 lexical resource for opinion mining, which assigns to each synset of WordNet [1] three sentiment scores: positivity, negativity and objectivity. After that, we calculated the accumulated sentiment of all messages, we color-coded the word in a green hue if the result is positive and above a configurable threshold, red if the result is negative and below a configurable threshold or gray otherwise. Restraining the seeds to account only for the profile's original posts will create a picture of their sentiment regarding a specific topic.

[1] WordNet is a lexical database for the English language. Synsets are sets of synonyms defined by WordNet. See https://wordnet.princeton.edu/

4.3 Bubble Chart

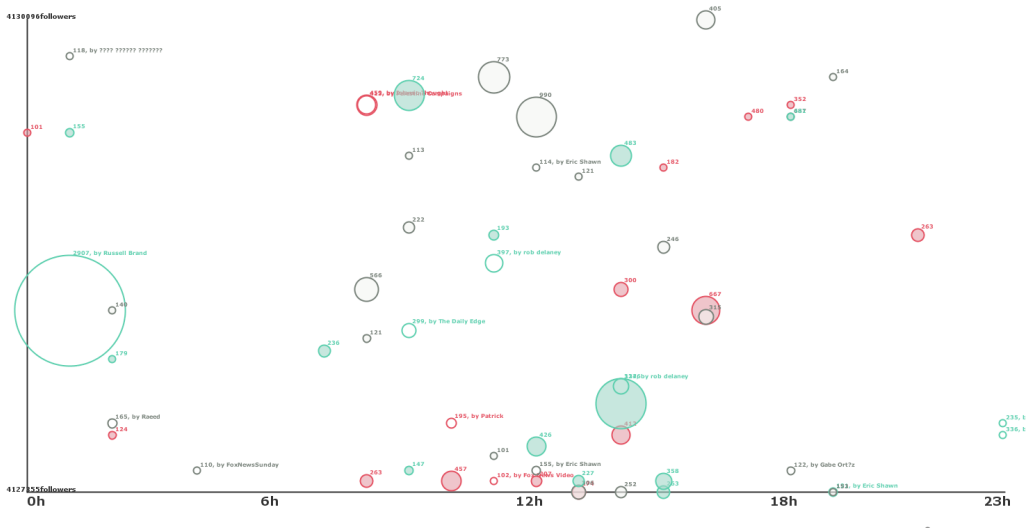


Figure 4.4: Bubble chart with re-tweeting activity color-coded by sentiment. Each bubble represents one original post by one of the selected sources. X axis represents time of day, ranging from 0 to 24, Y axis displays the number of absolute followers of all selected sources, from lowest to highest in the selected time span. Chart displays the time of the posting as well as popularity at the time, so the user can relate post traction to each dimension. A filled bubble means that the original post triggered the activity, while an empty bubble means that another user triggered the activity (middle-man). Selecting a bubble will filter the data set to only display the messages represented by that item in all other visualizations. Source: (BOETTCHER; COMBA; FREITAS, 2015)

The bubble chart (figure 4.4) displays the tweeting activity for a selected source and time period. The X axis maps the hour of the day when the message was posted and the Y axis represents the popularity the selected profile(s) had during the analyzed period, ranging from the smallest to the largest amount of followers.

The size of the bubble maps the number of re-tweets for the original message in the moment of the post. As part of the nature of the fire-hose API, we do not have guaranteed access to all the messages being posted at any time, but we can compare the number of re-tweets of the original post to get an idea of how popular each message became over time.

The bubbles are color-coded with the information as discussed in 3.2, with the added information of agent mapped to the fill of the bubble. Readers can comment or simply re-tweet messages (I) directly from the source, but the same activity may come from (II) a separate profile that mentioned, commented or re-tweeted the news channel. In figure 4.4, scenario (I) is represented by filled bubbles and (II) is mapped as hollow circles. In both cases, we include the raw number of re-tweets and in case

(II), we display the screen name of the profile that generated the activity.

4.4 Message Board

Figure 4.5: Message board color-coded by sentiment. Each message included in the selected and filtered data set is displayed on a board, ordered by descending date and time of the creation. Each message includes the author profile image, handle, date and time of the post and the message text. Selecting a message will filter out all messages that do not related to it from the other visualizations. Source: (BOETTCHER; COMBA; FREITAS, 2015)

Figure 4.5 shows the original message information that belongs to the selected sources and date and time being analyzed. As in the other techniques, sentiment is color coded as the background for each message. We included the original text, the date and time of the posting, the author's name, screen name and profile picture. This technique allows the user to get specific information in its lowest granularity and rawest form, which can help in identifying the context from which the tag cloud and bubble charts were derived.

Clicking on any message filters out every word or bubble from unrelated posts, enabling in-depth analysis of any conversation of interest.

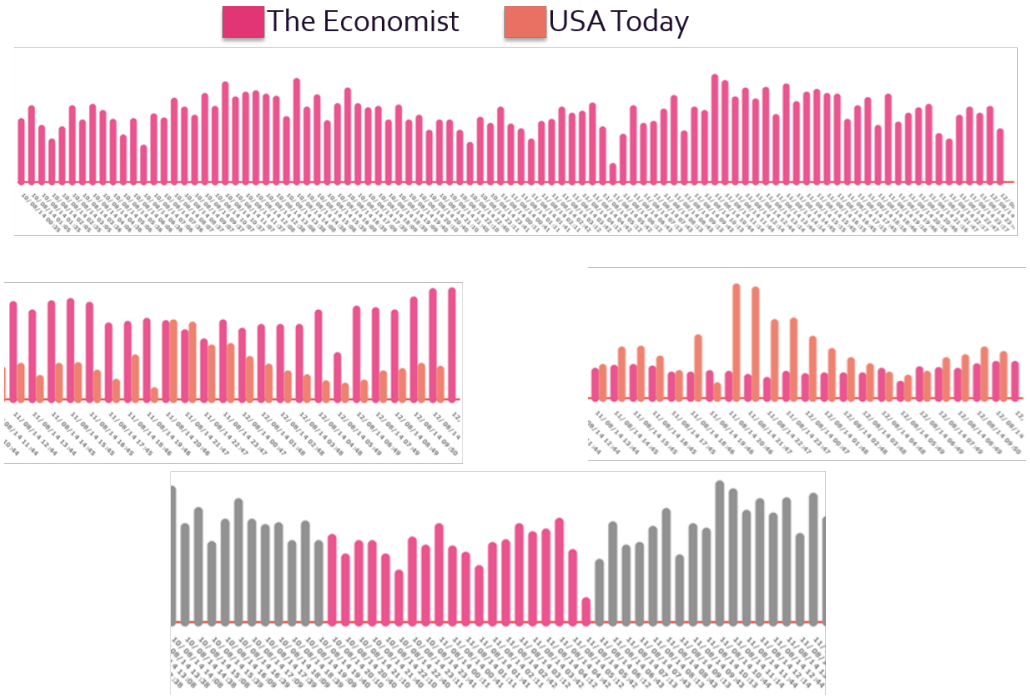


Figure 4.6: Bar charts plotting statistics of popularity based on the followers count of single (top, bottom) or multiple sources (middle tier) over time. Users can select an area of interest by dragging the mouse (bottom tier), which filters the data set in all visualizations. Users can create multiple windows of interest by holding the Ctrl (Cmd for Mac OS) key while dragging. Source: (BOETTCHER; COMBA; FREITAS, 2015)

4.5 Profile followers bar chart

The followers bar chart displays the patterns of popularity of the selected profiles according to the chosen metric over time as shown in figure 4.6.

This technique is very helpful in determining the relationship each news source has with their followers in terms of popularity. It allows the user to discover patterns in the growth of popularity of any channel and for the comparison between channels in a leveled playing field. This graph is a good starting point to discover points of interesting activity during an extended period of time. The user can select the appropriate window to restrain the information to the activity as shown in figure 4.6 (bottom) in order to research the window of interest and the remaining panels will filter out any information not comprised in that period.

This implementation is fundamental to create a popularity-based solution, and is not found in any other work of similar purpose.

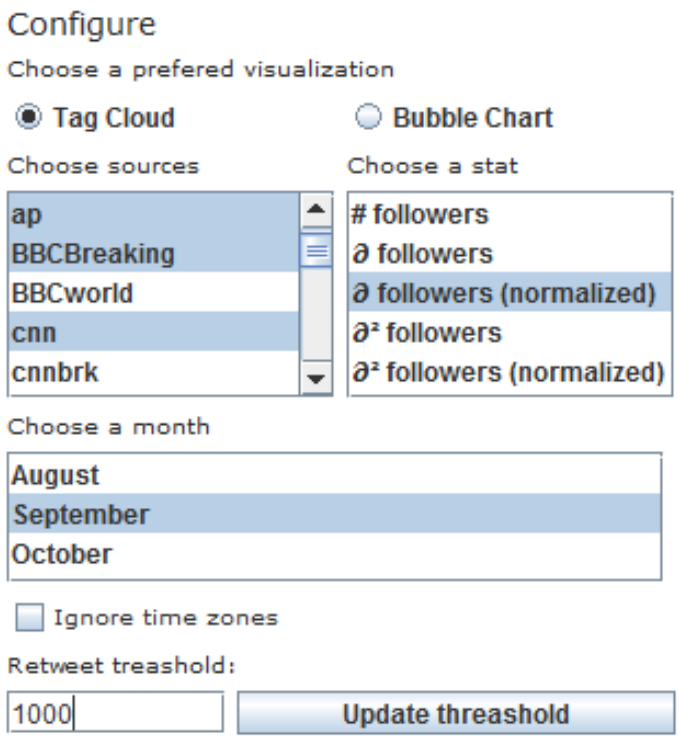


Figure 4.7: Configuration panel contains a choice of visualization (Tag Cloud and Bubble Chart), the list of available sources, five different statistics for the followers charts, the available monthly data sets, an option to ignore or consider time zones in the messages and a minimum threshold for the visualizations to filter out all posts that had fewer re-tweets. Source: (BOETTCHER; COMBA; FREITAS, 2015)

4.6 Configuration

The configuration panel is organized as in figure 4.7, where users can choose the visualization technique (Tag cloud and Bubble chart), and choose from the list of available profiles the ones they are interested in analyzing at each time. They can also choose from different metrics of behavior, i.e. total number of followers, difference in number of followers (delta) from each interval to the next, normalized delta from each interval to the next, delta squared and normalized delta squared of each interval. These two selections are necessary to display the bar chart of followers, shown in figure 4.6.

To optimize the search performance, we separated the data set into each month (August, September and October), so the user can easily switch between time windows of interest. The user also has the option of ignoring or taking the time zone into account while calculating the statistics. This separation is key to improve scalability.

The last item in the configuration panel is the re-tweet count threshold for the bubble chart. Only posts that surpass the minimum amount specified in that field will be displayed. This feature is particularly helpful when studying popular chan-

nels, such as CNN and the NY Times.

5 RESULTS

In order to determine the efficacy of our proposed solution, we invited the same journalists we interviewed in the beginning of the study to use the tool and use their expert knowledge for exploring the data to find answers for the four main questions that drove our research. We utilize key use cases to illustrate their reported findings and our analysis. We have already reported these results elsewhere (BOETTCHER; COMBA; FREITAS, 2015).

5.1 Use case 1

Comparison between different sources is not particularly key to understand the patterns of a profile's following base, but it can be very helpful in determining what other news sources have been doing and how that affected their popularity.

According to the journalists we interviewed, most of their activity in Twitter is automated, which means that information about their following crowd is not regarded while posting. When a new piece is included in the print version, it is automatically added in the channel's online publication after the editors finalize the next morning's issue. After the issue goes live, automatic processes create the messages and posts them automatically at the same time.

While worldwide news sources such as BBC and The Economist have a large follower count, the user is still able to compare them to local sources with a more limited reach, in order to understand if the community behavior follows a global trend or if their are driven by different motives.

As a filtering activity, the user selects the appropriate window to restrain the information to the activity as shown in figure 4.6 in order to look for some pattern pattern and select either the tag cloud or bubble chart visualization to better understand which were the causes of that behavior.

Three different views of the followers trends are shown in figures 5.1, 5.2 and 5.3. There are five selected profiles (Associated Press, BBC, CNN, The Economist and The Sun) based on three different metrics (Absolute number of followers, delta and

normalized delta).

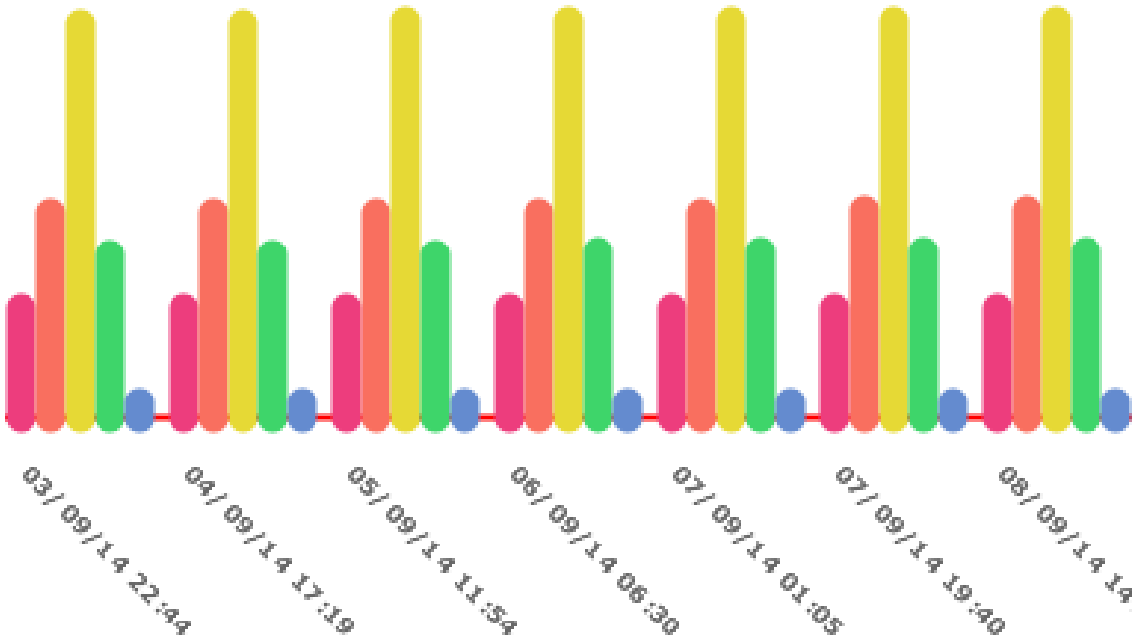


Figure 5.1: Bar chart of absolute number of followers for multiple sources of varying popularity. Each color represents a different Twitter profile: Associated Press in pink, BBC in orange, CNN in yellow, The Economist in green and The Sun in blue.

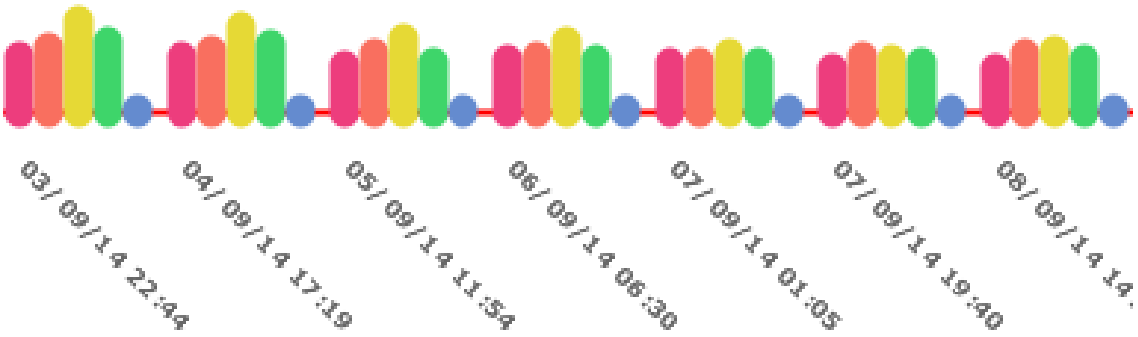


Figure 5.2: Bar chart of the delta of followers for multiple sources of varying popularity. Each color represents a different Twitter profile: Associated Press in pink, BBC in orange, CNN in yellow, The Economist in green and The Sun in blue.

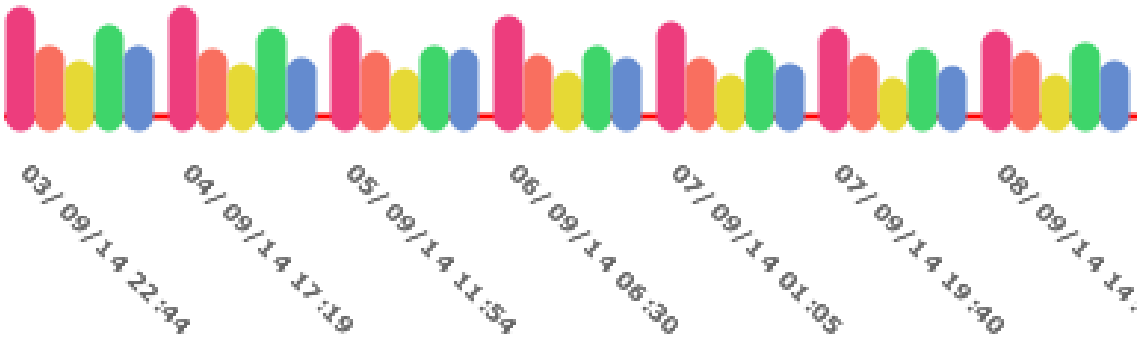


Figure 5.3: Bar chart of the normalized delta of followers for multiple sources of varying popularity. Each color represents a different Twitter profile: Associated Press in pink, BBC in orange, CNN in yellow, The Economist in green and The Sun in blue.

It is clear that CNN has a commanding lead in popularity, while The Sun has a very limited reach on Twitter, based on figure 5.1. It is also clear that the difference in popularity is sustained over time by each of the sources.

We can see in figure 5.2 that the number of followers increases by a similar amount for each of the profiles, The Sun being the sole exception. However, when we analyze figure 5.3, it becomes clear that the effect of growth is actually greater for the Associated Press profile. This analysis means that even though it may take a long time, the Associated Press and The Economist are diminishing the gap of popularity between them and BBC.

5.2 Use case 2

To understand the relationship between a given profile and its readers, we break their interaction into four aspects:

- Most common words used
- Time of original posts
- Time of re-tweets, comments and other related activity by the community
- Sentiments associated with each of the above

(a) Fox News only

(b) Fox News readers only

Figure 5.4: Comparison of original content provider and the reader community activity. Each tag is color-coded regarding the sentiment accumulated in the messages they appear. Green is used for positive tags, red is used for negative tags and gray is used for neutral tags.

During the analysis of the tool, we selected Fox News as the primary source to be analyzed and noticed the same pattern of popularity repeating itself over time, so we selected that window of dates, from August 9th to 11th. Figure 5.4a shows the most used words for Fox News in that time.

It is clear that the subjects of interest during this period for the profile were president Obama, Iraq, Robin Williams, ISIS and O'Reilly, which is not surprising since during that time, US jet fighters launched a strike on ISIS militants and Robin Williams passed away.

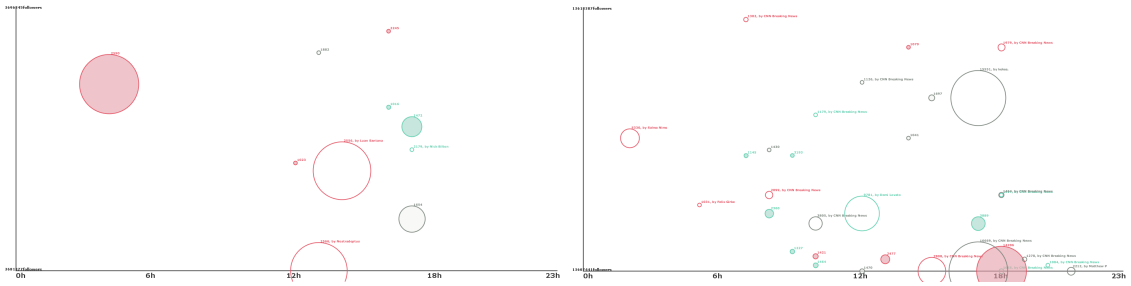
What is revealing is that even though Fox News is considered to have a republican bias, according to the analysis, they choose positive words when discussing president Obama, Iraq and the Islamics, while using negative words when mentioning their on-camera talent O'Reilly. Our users were expecting the opposite trend to be in place. Only the specific issues of "Obamacare" and Sen. Clinton are mentioned in a negative post, in confirmation to the journalists expectations.

By filtering the seeds to include the reading community's posts alone, the user is able to understand their opinions in the realm of the channels messages. Figure 5.4b shows the Fox News' community mentions the president in a contrasting neg-

ative fashion, while confirming the position on issues like ISIS. Another interesting contrast is the related profile @foxnewspolitics, which is mentioned positively by the main account but negatively by the community. This contrast poses a very interesting insight: even though Fox News may at times be favorable to the democratic issues, their community still mentions their content in a negative scenario.

In use case 3, we will describe further cases that relate to the time patterns found for other channels.

5.3 Use case 3



(a) Tweeting and re-tweeting habits of a local newspaper.

(b) Tweeting and re-tweeting habits of CNN.

Figure 5.5: Comparison of original content provider and the reader community activity. Each bubble is color-coded regarding the sentiment accumulated in the messages of re-tweet and reply to the original post. Green is used for positive tags, red is used for negative tags and gray is used for neutral tags.

The bubble chart technique allows us to understand how time and popularity influence the activity related to any given message as well as to identify when there are any key readers generating the interest instead of the original poster.

While the popularity of local and global news sources vary greatly, it is possible to identify similar trends in both communities: most of the activity takes place between noon and dusk. However, figure 5.5 shows how the local newspaper relies on separate influencer's re-tweeting or commenting of their post, most of the activity from the global news channel community is channeled through profiles from outside the institution. In contrast, the global source displays activity across the entire Y index, indicating that current popularity is not particularly key for generating large amounts of activity.

Figure 5.5a is an interesting case for the relationship of popularity and activity. Contrary to our initial estimates, activity from the community is inversely related to the popularity of the channel. After finding this pattern in our database in repeated occasions, we investigated the relationship of the source with the community via the

message board and discovered that this particular source is heavily dependent on key influencers to raise their popularity, even though this was not the case for actual activity related to their posts. Figure 5.5a also displays two significant influencers found in this local newspaper's community that will consistently (and significantly) increase activity around topics of interest, generating the greatest re-tweet counts of this particular source throughout the entire three months period. This specific discovery made by one journalist was deemed important to determine who are the key people that the publication can involve in order to increase its popularity.

5.4 Journalist reports

We organized a dedicated guided testing session of the tool with each of the journalists, which was comprised of a 20 minute tutorial of the visualizations and interactions followed by a 40 minute of assisted free play and a final questionnaire section to determine the tools efficacy and their impressions. Even though the majority of them found that performance was an issue, 75% reported that the interaction was very intuitive and provided a clear way to investigate any area of interest.

Most of the patterns found by the senior journalist were expected, based on his gathered knowledge over the years of work. However, the junior journalist was more intrigued by their findings. This points to the direction that the tool may be more beneficial to inexperienced professionals as a streamlined way to understand their public.

Based on these findings, there is an evidence that we have improved on existing information diffusion pattern seeking studies by enabling (1) the discovery of behavioral patterns that influence and explain the increase and decline of popularity of any specific channel, (2) the discovery of key third party influencers that actively shape the community's interests and (3) the comparison of similar and distinct sources' communities in a variety of ways.

6 CONCLUSIONS

Our solution utilizes known visualization techniques in innovative manners, aggregating information such as sentiment and popularity to give a unique view of the behavior of the community surrounding the news source profile and their relationship with each other.

Through the use of the proposed solution, it was possible to reveal hidden patterns and gain insightful knowledge of the reading community, which address real needs from the industry. The system improves the existing pool of solutions by using very clear parameters and established techniques to provide solutions to otherwise unanswered questions. While there are some gaps between the proposed motivations and our solution, based on the experts reports we understand that it shows considerable promise and believe that by enhancing the existing interaction, we will able to provide an important tool that can fine tune the way the press interacts with social networks.

The system produces convincing results for most scenarios but is not without its limitations. As mentioned in section 4.4, performance remains an issue especially for live data due to the heavy calculations needed to appropriately enable the interactions between panels. Complexity analysis on our heaviest algorithm, used for the calculation and rendering of the tag cloud visualization data, tends to $O(n^2 + k)$, where n is the number of posts selected for visualization and k is the number of sources selected to be displayed simultaneously. Like Zhao et al. (ZHAO et al., 2014a), this complexity poses issues with scalability. The added functionalities make the system more powerful, but also more complex. Depending on the unique patterns of a given community's activity, performance may also be affected negatively.

6.1 Future Work

We wish to continue developing our research by introducing methods for the user to analyze the reputation of different profiles with a large contribution to the

diffusion of the information. These techniques would improve on the information regarding the influential profiles, by allowing the user to understand the type of influence they effect on the community.

We would also like to expand form sentiment analysis to emotion analysis. Such evolution on the current algorithms would allow the users to gather key information regarding the effect a news source may have on the community as well as measure the success of marketing campaigns in social media.

Finally, we will look to improve on the time series to allow the automatic detection of temporal patterns and anomalies and suggest them as the initial focus points for the user.

6.2 Lessons learned

6.2.1 Data gathering

The process of data gathering for Twitter is limited by the API provided by the company, which limits the access to messages older than two weeks and requires a minimal amount of filtering. Additionally, the API limits the data flow to 30% of the posts only. Therefore, in order to have a complete data set, we recommend a well defined area of interest (by either subject or profiles) and the utilization of Fire Hose API to accumulate the tweets.

Each post object contains information on the author, including name and account location, message traction (e.g. re-tweet count, reply count) and geo-tagging information for each message (which may be different than the user account location). The most precise geographical information is the geo-tagging attribute in the message, however, we learned that only around 5% of users enable such information to be disclosed, so in order to compensate for the lack of information, we recommend building a local database of location names and its latitude and longitude correspondent. This way, the researcher can use the far more common account location information to infer a position for each of the messages.

Since the access to the original data is limited, we recommend tracking the amount of internal information contained in the JSON tags, such as number of profile followers, re-tweet count and location changes. These tags allow for the interpolation of the changes which in turn allow to understand how the community is changing its behavior patterns over time.

6.2.2 Interaction

The most success we experienced while building prototypes for the exploration of the social media data occurred when all the panels in the visualization tool reacted to any interaction by the user in any other section. When selecting any subset of data of interest in any visualization, users usually are positively surprised to see that other aspects of the tool will include their selection. Most users found these types of interaction helpful while exploring the information.

Since regular usage of any research tool entails the multiple filtering and selecting of areas of interest, we recommend the implementation of 'Undo' and 'Redo' functionalities, as well as a 'Reset' option in order to start the process from the beginning when needed.

6.2.3 Visualization

While the choice of visualization technique can be very particular to the data nature, there are a number of techniques that have been adopted by the majority of the users and that have been associated with typical representations. Using these techniques appropriately can ease the initial learning curve for the users, which in turn allows for a better understanding of how to fulfill their needs, either by adopting other well know techniques or by creating novel ones.

We first started introducing novel techniques to our users and the response was not as positive as we expected. Most users were having problems following the analogies we used to represent the patterns and to identify the meaning of the connections we were able to make. We followed the tests with improvements on the techniques, which took various interactions before we had a prototype that was ready for an actual analysis.

In our second research phase, we preferred the usage of more established techniques, which allowed us to both convey our intentions more easily to our audience and to make progress on more interesting aspects by gathering quality feedback from experts on the subject our data focused on without any background in computer science.

Appendices

AppendixA
EVALUATION STUDY QUESTIONNAIRE

Experiment The experiment consists in using the prototype to respond a few questions regarding your understanding of the dataset through the visualization and interaction.

1. Is there a difference in the diffusion of the news regarding the new Pope's election between Le Monde and The Economist? Use 'pope' as a filter.

2. Which location has the most re-tweets in this case?

3. Do you find any peculiarity in the flux lines in Europe when you use the word 'BRICs' as a filter?

Questionnaire

Thank you for your hard work so far. In order for us the measure the quality of the software effectively, we elaborated a questionnaire to gather your opinion after completing the tasks using the prototype. Please, answer the following questions marking the choice that most closely applies yo your opinion on the subject. The options are 5: 'Completely agree', 4: 'Partly agree', 3: 'Indecisive', 2: 'Partly disagree' and 1: 'Completely disagree'. Please select the last box ('N/A') if you have NOT utilized the feature mentioned.

Table A.1: Evaluation questionnaire

		5 4 3 2 1 N/A
1	The color scale representing the re-tweets count is adequate	□□□□□□
2	It is possible to easily identify the locations with the less activity	□□□□□□
3	It is possible to easily identify the locations with the most activity	□□□□□□
4	The colors used in the lines in the map facilitate the interpretation of the intensity of the location activity	□□□□□□
5	The connection between lines allows the temporal interpretation of the re-tweets with ease	□□□□□□
6	The divisions in the stream graph are interpreted with ease	□□□□□□
7	The colors of the divisions in the stream graph are interpreted with ease	□□□□□□
8	It is easy to compare two news source in the stream graph	□□□□□□
9	It is possible to mentally connect the stream graph and the map design	□□□□□□
10	It is easy to compare the behavior of the different locations in regards of a given subject	□□□□□□
11	It is easy to detect dependent behavior patterns in the days of the week in the stream graph	□□□□□□
12	The legends in the stream graph are informative	□□□□□□

REFERENCES

ANGEL, A. et al. What's on the grapevine? In: **INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA SIGMOD, 2009. Proceedings**. New York, NY, USA: ACM, 2009. p. 1047–1050.

BOETTCHER, G. B.; COMBA, J.; FREITAS, C. D. S. News patterns: how press interacts with social networks. In: **WSCG'2015 - 23rd INTERNATIONAL CONFERENCE IN CENTRAL EUROPE ON COMPUTER GRAPHICS, VISUALIZATION AND COMPUTER VISION, 2015. Proceedings...** [S.l.: s.n.], 2015. p. 143–150.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. In: **SEVENTH INTERNATIONAL WORLD-WIDE WEB CONFERENCE, WWW, 1998. Proceedings**. [S.l.: s.n.], 1998.

CHAE, J. et al. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: **CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY, VAST, 2012. Proceedings...** [S.l.]: IEEE, 2012. p. 143–152.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 41, n. 3, p. 15:1–15:58, jul. 2009. ISSN 0360-0300.

DEERWESTER, S. C. et al. Indexing by latent semantic analysis. **JASIS**, p. 391–407, 1990.

HAKLAY, M. M.; WEBER, P. Openstreetmap: User-generated street maps. **IEEE Pervasive Computing**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 7, n. 4, p. 12–18, oct. 2008. ISSN 1536-1268.

HAVRE, S.; HETZLER, B.; NOWELL, L. Themeriver: visualizing theme changes over time. In: **SYMPOSIUM ON INFORMATION VISUALIZATION**. [S.l.]: IEEE, 2000.

KWAK, H. et al. What is twitter, a social network or a news media? In: **WWW Proceedings**. [S.l.: s.n.], 2010. p. 591–600.

LERMAN, K.; GHOSH, R. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: **ICWSM**. [S.l.: s.n.], 2010.

LU, Y. et al. Integrating predictive analytics and social media. In: **CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY (VAST), VAST, 2014**. [S.l.]: IEEE, 2014.

MATHIOUDAKIS, M.; KOUDAS, N. Twittermonitor: trend detection over the twitter stream. In: **INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, SIGMOD 2010. Proceedings...** New York, NY, USA: ACM, 2010. p. 1155–1158.

PLUTCHIK, R. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. **American Scientist**, Sigma Xi Scientific Research Society, v. 89, n. 4, p. 344–350, 2001.

SUN, G. et al. Evoriver: Visual analysis of topic coopetition on social media. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 1753–1762, Dec 2014. ISSN 1077-2626.

TURNER, J. D.; ELGOHARY, T. A simple perturbation algorithm for inverting the cartesian to geodetic transformation. In: **Mathematical Problems in Engineering**. [S.l.: s.n.], 2013.

VIEGAS, F.; WATTENBERG, M.; FEINBERG, J. Participatory visualization with wordle. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1137–1144, Nov 2009. ISSN 1077-2626.

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**. [S.l.]: Cambridge university press, 1994.

WU, Y. et al. Opinionflow: Visual analysis of opinion diffusion on social media. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 1763–1772, Dec 2014. ISSN 1077-2626.

YANG, J.; COUNTS, S. Comparing information diffusion structure in weblogs and microblogs. In: **ICWSM'10**. [S.l.: s.n.], 2010.

YANG, J.; COUNTS, S. Predicting the speed, scale, and range of information diffusion in twitter. In: **ICWSM**. [S.l.: s.n.], 2010.

YE, S.; WU, S. F. Measuring message propagation and social influence on twitter.com. In: **SECOND INTERNATIONAL CONFERENCE ON SOCIAL INFORMATICS, SOCINFO, 2010. Proceedings...** Berlin, Heidelberg: Springer-Verlag, 2010. p. 216–231.

ZHAO, J. et al. Fluxflow: Visual analysis of anomalous information spreading on social media. **, IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 1773–1782, Dec 2014. ISSN 1077-2626.

ZHAO, J. et al. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In: **CONFERENCE ON VISUAL ANALYTICS SCIENCE AND TECHNOLOGY (VAST), VAST, 2014. Proceedings...** [S.l.]: IEEE, 2014. p. 203–212.