UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

SPATIO-TEMPORAL HISTORICAL EVENT VISUAL EXPLORATION THROUGH SOCIAL MEDIA-BASED MODELS

TESIS PARA OPTAR AL GRADO DE DOCTORA EN CIENCIAS, MENCIÓN COMPUTACIÓN

VANESSA CAROLINA PEÑA ARAYA

PROFESOR GUÍA:
BÁRBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
BENJAMÍN BUSTOS CÁRDENAS
NANCY HITSCHFELD KAHLER
TOBIAS SCHRECK

Este trabajo ha sido parcialmente financiado por CONICYT,
INSTITUTO MILENIO EN FUNDAMENTOS DE LOS DATOS

SANTIAGO DE CHILE
2018

Resumen

Las plataformas de redes sociales en línea sirven como importantes fuentes de información acerca de lo que está pasando en el mundo y cómo la gente reacciona a estos eventos. Dentro de toda la información útil que los científicos han extraído de estos repositorios, el análisis de mensajes relacionados con eventos del mundo real son una importante oportunidad para realizar análisis histórico de noticias. Como los mensajes publicados en estas plataformas contienen distintos puntos de vista de una noticias, contribuyen con información que quizás no haya sido publicada por los medios tradicionales. Dentro de los aspectos que se pueden estudiar de un evento noticioso, las relaciones geopolíticas como consecuencia de ellos contienen información valiosa para análisis histórico futuro. En efecto, entender las relaciones entre países, su desarrollo en el tiempo y cómo las personas reaccionaron a ellas es esencial para comprender el presente.

Sin embargo, extraer información útil desde estas plataformas no es una tarea fácil dada la creciente velocidad de publicación de sus mensajes, lo no estruturado de su contenido y la enormidad de repositorios que generan. Por otra parte, para extraer conocimiento nuevo se necesitan herramientas que permitan la generación de hipótesis nuevas por parte de expertos en un dominio. Esta necesidad de colaboración entre sistemas computacionales y usuarios finales hace que el problema tenga dos componentes. El primer componente es que los datos pueden ser difíciles de guardar, recuperar y procesar sin las representaciones adecuadas de alto nivel. El segundo componente es que explorar con ojos humanos un gran número de mensajes puede ser imposible sin las herramientas adecuadas.

El objetivo de esta tesis es abordar estos dos problemas. El primer problema, relacionado con la eficiencia del procesamiento computacional de los datos, se aborda presentando una representación de alto nivel de eventos noticiosos en su contexto geopolítico. Más específicamente, proponemos una *representación de eventos consciente del contexto espacio temporal* que incorpora tanto la información de las ubicaciones que están involucradas en el mundo real como de aquellas hasta donde se propagó el evento. Exploramos la utilidad de este modelo usando datos de eventos noticiosos extraídos desde Twitter en una ventana de tiempo de dos años. Abordamos el segundo problema, relacionado con la exploración de mensajes por expertos en un dominio, diseñando herramientas visuales para exporarlas. Primero diseñamos una interfaz web visual, llamada *Galean*, que permite a usuarios explorar noticias dada la representación de eventos anteriormente mencionada. Evaluamos esta interfaz a través de un estudio cualitativo con potenciales usuarios finales y uno cuantitativo con 30 participantes. Dada la retroalimentación recibida en esas instancias, diseñamos y evaluamos una nueva manera de visualizar datos geográficos y temporales llamada *Cartoglyphs*.

Abstract

Online social media platforms serve as important information resources on what is happening in the world and how people react to these events. Among the valuable information that scientists have extracted from these repositories, analyzing the messages related to real-world events presents an important opportunity for historical analysis of news. As messages published in these platforms contain diverse points of view of a news event, they contribute with information that might not have been brought to the public by traditional news media. Among all the angles from which a news event can be analyzed, the geopolitical interactions among countries as consequence of news events contain valuable information for future historical analysis. Indeed, understanding the relationships among countries, their development over time and people's reaction to them is essential to comprehend the present.

However, it is not easy to extract useful information from the messages published in social media platforms. This is because these messages come at an increasingly fast rate, their content is unstructured, and they make up huge data repositories. On the other hand, extracting new knowledge from this data requires tools that allow domain experts to generate new hypotheses. This need of collaboration between computational systems and final users makes the problem two folded; on one hand, the data can be computationally difficult to store, retrieve and process without proper high level representations; on the other hand, manually exploring a large number of text messages can be unfeasible without proper tools.

The objective of this dissertation addresses these two problems. The first problem, relating to the computational effectiveness of processing the data, is addressed by presenting a high level representation of news events by their geopolitical context. More specifically, we propose a *spatio-temporal context-aware event representation* that incorporates information about the locations that are involved in the real-world and the locations to where an event was propagated. We then explore the usefulness of this model using two years of data of news events collected from Twitter. We address the second problem of exploration of messages by domain experts through the design of visual tools. For this component, we first designed a visual web interface that allows users to explore news events given the proposed model which we call *Galean*. We evaluated this interface by conducting a qualitative study with potential final users and a quantitative one with 30 participants. Given the feedback received in those instances, we designed and evaluated a new way to visualize geographical and temporal data called *Cartoglyphs*.

To my mom, Paulina and Viviana:
my three favorite people.

Para mi mamá, Paulina y Viviana:
mis tres personas favoritas.

Acknowledgements

Oh, gods, it has been such a long long trip. I do not think I will do a PhD again, thank you very much. I would love to thank so many people that were part of this process. However, since my native language is Spanish it is a bit uncomfortable to write the most sentimental part of this document in English. For this reason I'll switch to Spanish for a while and then return to the normal English.

Como usualmente se parte en estas instancias, quisiera agradecer a los profesores que me acompañaron desde que decidí seguir este camino. Primero que nada quisiera agradecer a Bárbara Poblete, mi profesora guía, por su confianza en mi trabajo y su apoyo en el desarrollo de esta tesis. También quisiera agradecer a Alexandre Bergel, quien me motivó a trabajar en el área de visualización y quien fue uno de los motivadores principales para que hiciera el doctorado. Gracias también a la profe Nancy Hitschfeld por sus comentarios acerca de mi trabajo, este y anteriores, y por la colaboración en los proyectos de Scratch. Muchas gracias por todo lo que me enseñaron en estos años, por su (mucha) paciencia y buena onda. Mil gracias también a Benjamín Bustos, Denis Parra y Cecilia Aragon, profesores con los que también tuve la fortuna de trabajar durante mi doctorado.

Quisiera a CONICYT, por financiar mis estudios de doctorado a travs de la beca CONICYT-PCHA/Doctorado Nacional 2013/21130470. También quisiera agradecer el financiamiento parcial del Instituto Milenio en Fundamentos de los Datos, financiado con fondos de la iniciativa Milenio, y programa de Becas NIC Chile.

Infinitas gracias a Jazmine y a todo el equipo de Niñas Pro, por su dedicación, por ser tan movidas y hacer que este proyecto maravilloso crezca cada día más.

Gracias a Jorge, Jorge padre, tía María Eugenia, abuelita Luz, tío Pato, Cecilia, Juan Pablo, Gus y Magda, por acompañarme muchos años y acogerme como parte de su familia. A la Mabel, por existir, ser adorable e inteligente. A todos los amigos que conocí en la universidad, principalmente a Daniel, Carlos, Lis y a la Conejo. A Jérémy por su amistad, por sus consejos y por ayudarme a corregir las primeras versiones de este documento. A Pavlov por las juntas a copuchar, llorar y celebrar. A Juampicillo, por su compañía estos años, los chistes fomes y los viajes a Bolivia.

Gracias a mis compañeros de oficina, Mauricio y José Miguel, por no venir muy seguido y dejarme la oficina para mí sola. Gracias a Renato Cerro por todas sus correcciones de inglés y las discusiones filosóficas acerca de la diferencia entre datos, información y conocimiento. Gracias a sandra y Angélica por ayudarme en todo lo administrativo pero también por

escucharme y aconsejarme en los momentos que lo necesité.

Gracias al Migue por Cocha, por las sopaipillas más fomes del mundo, por la lista interminable de canciones, por julio y por los Cigarreins.

Hay amigas que aportaron absolutamente nada en esta tesis, de hecho ni siquiera las conocía cuando empecé a escribir el borrador de este documento, pero es importantísimo que las mencione. Primero que nada a Salomé, la francesa más choriza y seca que conozco. Gracias por enseñarme a subir montañas sin morir y a usar bananos, el accesorio de vestuario más horrendo del mundo y el más cómodo. A Laura, que me enseña a cantar cuando logramos no ponernos a conversar comiendo sushi y tomando vino. A Érika, por compartir tatuadores y por buscar donde viviremos en Francia. Gracias a las tres por escucharme, por sus abrazos, por su cariño infinito y por contribuir a no volverme loca.

Finalmente, gracias a las tres personas que más amo en el mundo. Gracias a mi mamá por su sabiduría, por creerme y por apoyarme tanto en estos últimos años. A la Pauli, por sus recomendaciones de pelis, por los bailes de cueca y por su cariño. A la Vivi, por las sesiones de psicóloga gratis por Whatsapp y por hacer el ridículo conmigo en el metro.

Now, we are back to English and to the thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Internet, beyond connecting computers, connects people: It started with simple text messages through email, and now continue with a huge variety of tools and platforms. In particular for this dissertation, we consider social media platforms, where a message published by a user can reach several other users at once. This type of connection among people allows information to be rapidly shared. In this work, we use the definition of social media of Kaplan et al. [93]:

> *Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.*

There are several social media platforms available, such as Facebook [1], Youtube [2] or Instagram [3]. Twitter [4] is one of such platforms, ranked number 12 in the most visited websites worldwide in Alexa's web traffic analysis [10]. Twitter is defined as a micro-blogging service in which users can create an account and publish messages called *tweets*. Initially, tweets were limited to be only of 140-characters long. However, this constraint was changed on November 2017, now allowing 280 characters. When users publish a message, it will be broadcast to all their *followers* (i.e., the users that are subscribed to the account), who can read it, reply to it or broadcast it further to their own followers. Twitter users mainly use this platform as an information source instead of a platform for satisfying social needs [105].

The fact that Twitter users utilize this platform as an information source and an information broadcaster is reflected in the behavior of users: when a breaking news event occurs they quickly react by generating content and producing interactions. This reaction facilitates rapid exchanges in the social network, allowing the quick propagation of news. Indeed, as expressed by the Twitter web page, usage spikes when something relevant happens in the world [5].

[1] `https://facebook.com`
[2] `https://youtube.com`
[3] `https://www.instagram.com/`
[4] `http://twitter.com`

In particular, the public access to the platform allows users to disseminate information without necessarily passing through an editorial review or censorship, in contrast to traditional media. This concept is commonly known as citizen journalism [101]:

> *Citizen journalism is any type of journalism engaged in by someone who has not undergone formal training to be a journalist and, in most cases, is not subject to oversight or censorship.*

As consequence, with these platforms, particularly Twitter, control over information is more decentralized and democratic [101]. Furthermore, in some occasions the information published in Twitter can influence how traditional media builds its agenda [139].

The need for means to make sense of this data is reflected in the increasing body of scientific work directed at understanding social media information. For instance, Liu et al. [117] used Twitter to create a system that detects real world news events in real time and computes a score to predict the veracity of the event. As the content published in Twitter does not pass through any filter, there are several investigations on the credibility of content propagated on this platform [33, 79, 80]. In addition, researchers have analyzed protests and other social movements by extracting messages from this platform. Some examples are the work of Scherman et al. [161], the one from Eltantawy et al. [56] or the one from Blandford et al. [25].

Geographical features are an important aspect of news. In particular regarding social media platforms, there are several studies available in the literature about the geographical propagation of content in Twitter [86, 6, 204]. However, to the best of our knowledge none have focused on the political relationships among countries that can be extracted in the messages published in these platforms. The study of the geopolitical characteristics of an event allows us to understand present tensions or alliances among countries that influence news. Furthermore, observing geo-temporal patterns of behavior allows researchers to study news events in depth and predict future relationships among locations, for example. Social media contains geopolitical information that allows this type of analysis. For example, the tweet *"U.S. warns #NorthKorea against new missile test, plays down talks - and reports about how they may be planning another missile test soon"* published on January 2nd of 2018, suggests a political tension between the United States of America and North Korea given a news event in the physical world. To understand this event in its historical context, a person might like to explore past events that describe the relationship between both nations and to follow its evolution to understand the present. They might also like to explore people's reactions in social media and the impact it had in other places of the world. Given this inquiry, this thesis aims to answer the following main question: *is it possible to extract meaningful information about the geopolitical interaction of countries given news events when analyzing social media data on a large scale?* Given the extremely large volume of the messages coming from the Twitter streaming, this data becomes very volatile and difficult to analyze. To answer that question this thesis focuses on researching how to efficiently visualize social media data in ways that allow for geopolitical analysis of news events, in addition to the identification of geo-temporal patterns of people's and news.

1.1 Problem Statement

The extremely large volume of streaming messages makes information on Twitter very volatile. In addition, given their unstructured nature it is extremely difficult to gain high level insight about events, even less about a set of events over time. On the other hand, understanding patterns in news behavior is impossible without human analysis along with computer data processing. Indeed, even with high level Data Mining tools or methodologies, it might not be possible to analyze the messages commenting on a news event in abstract ways as humans do. We therefore divide the problem of extracting meaningful data in two main components:

- The data can be computationally difficult to store, retrieve and process without proper high level representations. In particular, it can be hard to represent Twitter data to include the geopolitical entities involved in news events in addition to the geographical places to where they were propagated. Furthermore, it can be a difficult task to contrast both views.

- Manually exploring a large number of text messages can be unfeasible without proper tools. Moreover, when analyzing more than one aspect of a news event, it can be difficult to visualize multiple variables at once. Hence, visualizing geo-temporal change is not an easy task.

We believe in the importance of finding solutions to these problems in order to contribute to the historical registry of virtual documents published in social media.

Work Hypotheses and Research Questions

We work under two main hypotheses with their sub-hypotheses:

- **H1:** The data published in social media platforms contains valuable information about what is happening in the real-world.
  - **H1.1**: Analyzing data from social media yields historical data about news related to geopolitical interaction among countries as consequence of news events.
  - **H1.2**: By analyzing social media data one can understand how people reacted to a news event and the geographical places to which news events propagated to.
  - **H1.3**: By analyzing data from social media one can obtain insight of how events relate to each other over time.
- **H2**: Visual representation of news in their geopolitical context allows users to extract valuable knowledge about the real world.
  - **H2.1**: An expressive visual representation allows users to visually identify and extract patterns, which cannot be easily found through manual or quantitative

analysis of the raw data.

- **H2.2**: An expressive visualization of the geopolitical context of a news event allows users to extract relationships among events and participating entities.
- **H2.3**: A simple visual representation of geographical data allows users to extract knowledge from several points of view of a news event.

The research questions we aim to answer are:

1. Which geopolitical entities participate in real-world events?
2. How do these geopolitical entities interact as consequence of real-world events?
3. How do these geopolitical relationships evolve over time?
4. How do real-world events propagate around the world through social media?
5. What is the reaction of people in social media when a particular geopolitical entity gets involved in a particular real-world event?
6. What features of a real-world event are important to model in order to retrieve it later in an effective way?
7. Which visual features allow users to understand a news event in their geopolitical context and its evolution over time?
8. When considering journalists as final users to conduct this type of analysis, which features are important in a visual interface supporting it?

1.2 Objectives

The main goal of this thesis is to design a visual representation, and an underlying model to support it, for news events extracted from social media. The visual interface should allow users to explore, search, retrieve and analyze news events in their real world characteristics (the event itself) and their impact on people (social media) through their geopolitical and temporal context. We extract news events from Twitter's data repository. The detected news events are defined as a real-world occurrence that are reflected in social media and detected by given a methodology. It is important to note that we work with the existing literature, though our contribution is not about how to detect events from social media. More specifically, we divide our general objectives (labeled with numbers) and specific objectives (labeled with letters) as follows:

1. Related to **H1**: Study, model and analyze the geopolitical context of news events detected from online social networks and their evolution over time.
   (a) Create a high-level contextual event model that considers the geopolitical rela-

tionships among locations, their impact in social media and change over time.

   (b) Perform quantitative and qualitative analysis of the proposed context-aware models in order to study similarities among events and other interesting patterns that may emerge over time.

   (c) Study non-trivial information that can be automatically inferred about the physical world by automatically analyzing the proposed context-aware models, for example: information about international relations and the impact that an event had on people.

2. Related to **H2**: Create visualization tools that support using the proposed model described above.

   (a) Identify an effective visual representation of news events that allows users to understand such events by the countries involved, the countries participating in the social debate about them, their relations and development over time;

   (b) Study the user domain (e.g., journalists) needs and feedback for the visualization tool.

1.3 Methodology

Given the objectives described previously, we define the methodology as follows.

Methodology for General Objective 1 and Its Specific Objectives:

1. Design a location-aware model for news events using social media data. This model contemplates the location involved in the physical-world event, the locations to where the event was propagated in social media, the time of the event, etc.

2. Generate this model using a collection of news events generated using an external methodology of event detection on social media. This data was enriched by extracting the geographical data present in the text of the messages discussing the event and the location of the users who published them.

3. Use an exploratory data mining based approach to identify interesting patterns in the event collections, such as: sets of similar events, similar countries according to their participation in common events in time, event tracking in time, etc.

Methodology for General Objective 2 and Its Specific Objectives:

1. Design and implement a Web based visualization tool that allows for the exploration of events using social media data. Using the model and data described in the previous step, the user will be able to retrieve events by date, locations involved, relevant keywords and/or its scope.

2. Evaluate the initial design using qualitative and quantitative approaches to detect problems and extract relevant feedback from potential final users.

3. Iterate the design of the visualization given the user studies.

1.4 Contributions

The contributions of this thesis dissertation are summarized as follows:

- *A news event representation* - to analyze events in their geopolitical and temporal context we present a news event representation that considers the provenance of an event and its impact in social media.

- *Data insight* - by using the event representation presented above, we explore a 2-year database of news events and visualize relationships among countries and events.

- *Galean* - A visual interface for representing events by the above mentioned model. In addition, we also describe two user studies with feedback and comments about the interface.

- *Cartoglyphs* - A new way of representing geographical data as cartograms reduced as glyphs. We present initial designs and an empirical study of their usefulness.

As result of the work conducted in the present dissertation, the following articles have been published:

- **Vanessa Peña-Araya,** Mauricio Quezada, Barbara Poblete and Denis Parra, Gaining Historical and International Relations Insight from Social Media: A Spatio-Temporal Context-Aware Model for News Events in Twitter. EPJ Data Science 6.1 (2017), 25

  URL: `https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0122-8`
  Number of pages: 35 (full paper)

- Mauricio Quezada, **Vanessa Peña-Araya** and Barbara Poblete, Location-Aware Model for News Events in Social Media, SIGIR 2015, pages 935-938.

  URL: `https://dl.acm.org/citation.cfm?id=2767815`
  Number of pages: 4 (short paper)

- **Vanessa Peña-Araya**, Mauricio Quezada and Barbara Poblete, Galean: Visualization of Geolocated News Events from Social Media, SIGIR 2015, pages 1041-1042.

  URL: `https://dl.acm.org/citation.cfm?id=2767862`
  Number of pages: 2 (demo paper)

- Jazmine Maldonado, **Vanessa Peña-Araya** and Barbara Poblete, Spatio and Temporal Characterization of Chilean News Events in Social Media. SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access (TAIA15), August 13, 2015, Santiago, Chile.

  URL: `http://research.microsoft.com/en-US/people/milads/taia15-5.pdf`
  Number of pages: 4 (workshop paper)

- **Vanessa Peña-Araya**, Jorge Bahamonde, Barbara Poblete and Benjamin Bustos: Cartoglyphs: Reducing the World to a Glyph for Quick Exploration and Comparison of Spatio-Temporal Change. Poster session at IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 2016.

  URL: `http://ieeevis.org/year/2016/info/overview-amp-topics/posters`
  Number of pages: 2 (poster paper)

Papers currently under revision:

- **Vanessa Peña-Araya,** Barbara Poblete and Benjamin Bustos, Cartoglyphs: Studying the Use of Glyphs to Visualize Geo-Temporal Evolution Over Time. We are currently working on a revised version to submit to the Sage Information Visualization journal.

1.5 Thesis Summary

This dissertation is structured as follows:

- **Chapter 2: Background**, in which we describe relevant concepts for the present thesis.

- **Chapter 3: Literature Review**, describes research work relevant to this work, related to event modeling and geo-temporal visualization.

- **Chapter 4: Geo-temporal Representation of Events Extracted from Social Media**, details the event definition for geo-temporal analysis and the exploratory analysis conducted with it.

- **Chapter 5: Visualization of News Events by their Geo-temporal Representation**, the prototype of the visual interface that implements the proposed event representation is presented, including details about its system architecture, its interface components and its validation.

- **Chapter 6: Cartoglyphs: Visualizing Geographical and Geo-temporal Data with Glyphs**, describes the proposed new visualization to display geographical data as glyphs for a simplified view of the world. It also contains the details about the conducted evaluation of proposed glyphs design.

Chapter 2

Background

In this section we describe important concepts that easily explain the present dissertation. In particular, we focus on the following areas:

- Twitter, mainly as this social media platform is used as the data source of the news events used for analysis.
- Document representations in order to understand the methodology to extract new events from Twitter and related literature.
- Geographical data related concepts, as this dissertation focus on the spatio-temporal aspects of news events.

2.1 Twitter

Twitter is an online social media service in which users can create a profile and publish messages called *tweets* in their *timeline*. The timeline of a user contains all the messages published by him or her. These messages can be public or private, depending on the user preferences. Twitter users connect to each other by *following* other users in asymmetrical relationship: a user can follow another without their reciprocity. If a user's account is public, all the messages published in his timeline can be read by his followers. There are several characteristics of Twitter that differentiate from other social media services:

- Twitter messages were originally constrained to 140 characters. As of November 2017, the limit is now 280 characters.
- Each message contains medatada, such as the date when it was published, the user who did it, the location of the user who published it, among others. In addition, a tweet can contain media such as videos or images.
- A tweet can be a *reply* to another tweet, usually containing the user or users who

participate in the conversation.

- A tweet can be a *retweet*, which is a report or forward of a message published by another user.
- A user can *mention* another user by using `@username`, which is notified to the mentioned user.
- Tweets can contain *hashtags*, which are words or phrases used to represent a topic. A tag is composed by a # as prefix and then the word or phrase, like *#BlackLivesMatter*.

To collect tweets, Twitter provides an API [178], which provides developers a way to get a sample of tweets in real time. The returned sample is a percentage of the tweets publicly available that were recently posted. Another way is to search for tweets with certain characteristics, such as those containing particular keywords or that they were published from a bounded geographical area. A tweet message is described by 31 features, a summary of a selected set of them is in Table 2.1. The data about the user who published a tweet is composed of 37 data attributes, from which a subset of them are described in Table 2.2.

Table 2.1: Subset of relevant features available for a Twitter message

Feature(s) name(s)	**Description**
id	A unique identifier which can be used to access to a tweet via url, as in `https://twitter.com/user_name/status/id`.
text	The actual content of the message.
coordinates	If available, the geographical coordinates from where a tweet was published
user_id	The unique identifier of the user who produced the tweet
lang	The language in which a tweet is published.
created_at	The date and time in which a tweet was published.

2.2 Document Representation

To contextualize the extraction of topics or events from data streams, in this section we describe some of the most commonly used models to represent documents and related concepts.

Text documents are not immediately readable by computers because natural language is complex and diverse. Because of this, the transformation from a human readable format to a structured model that a computer can interpret requires preprocessing the text of the document. One of the first steps in this transformation is *tokenization*, which is the process of cutting the text into words or sentences called tokens. Sometimes at this step some characters can be removed, such as punctuation or stop words. Later, to group words that contain similar information, there are processes like *stemming* and *lemmatization*. Stemming

Table 2.2: Subset of relevant features available for a Twitter user

Feature(s) name(s)	**Description**
user_id	The unique identifier of an user.
name	The name of the user by their own specification.
verified	A verified Twitter account will be marked with a official blue verified tick badge. This verification is usually sought by those who are at risk of being impersonated, such as celebrities or public figures. This field indicates whether a user has been verified or not.
geo_enabled	Indicates whether or not the user had indicated that her or his tweet can be geotagged.
location	The location the user provided in text, if available.
protected	Indicates whether or not the user decided to protect their tweets, in which case they are not publicly available.

is the process of reducing a word to its root. For example, the words fishing and fisher could be reduced to fish. However, this process does not always yield correct answers. For instance, jumping and jumpiness could be reduced to jumpi instead of jump. On the other hand, lemmatization is the process of determining the lemma of a word based on its intended meaning. The words paying, paid, and pays will be lemmatized to pay, which is their lemma.

After a document has been preprocessed, it is possible to convert it to a selected model. One of them is the bags-of-words model, which represents a document by its words and their frequency. For example, the sentence "The quick brown fox jumps over the lazy fox", can be represented as:

$$\mathrm{d} = \{\text{“}the\text{”} : 2, \text{“}quick\text{”} : 1, \text{“}brown\text{”} : 1, \text{“}fox\text{”} : 2, \text{“}jumps\text{”} : 1, \text{“}over\text{”} : 1, \text{“}lazy\text{”} : 1\}.$$

It is important to note that this model does not focus on the order or the semantics of the words. The n-gram model is similar to the bag-of-words model, but instead of considering the frequency of each word, it considers the frequency of n consecutive words on a document. The previous example could be converted to n-grams like:

$$\begin{aligned}\mathrm{d} = \{&\text{“}The\ quick\text{”}, \text{“}quick\ brown\text{”}, \text{“}brown\ fox\text{”}, \text{“}fox\ jumps\text{”},\\ &\text{“}jumps\ over\text{”}, \text{“}over\ the\text{”}, \text{‘}the\ lazy\text{”}, \text{“}lazy\ fox\text{”}\}\end{aligned}$$

Then, the probability of occurrence is computed for each n-gram. For instance, in a document written in English it will be more likely to find the sentence "good morning" than "morning good", so the first one should have a higher probability of occurrence than the second one. In this sense, it stores part of the context in which some words appear together,

allowing the estimation of the occurrence probability of a given sentence. These two models are related, as the bag-of-words can be thought as a particular case of the n-gram model in which $n = 1$. The intuition for using these models for clustering or classifying is that similar documents have similar content. For example, if we compare the bags of words produced by each the books of The Lord of The Rings saga, they should have similar vocabulary among them. On the contrary, if we compare them with those of the Twilight novel series they should be different.

Another document representation is the vector space model, which is an algebraic model used to represent documents as vectors. In particular, given a bag of words, it is possible to create a vector space representation of the document in which each word is a feature. More specifically, each word in the vocabulary represents an axis which defines a $|V|$-dimensional space. In this space, each document is positioned in this space given the weight associated to each word, which can be computed by a function. The intuition of this model is that similar documents will be closer in the space of the vocabulary. Formally, a document in this model can be defined as:

$$\mathrm{d} = (w_{t1}, w_{t2}, ..., w_{t\|V\|})$$

The function that computes the weight of a term (w_{i}) in a document can be one that returns a binary value indicating whether is present in the document or not, or one that returns the frequency of its appearance. The frequency of a term indicates how important a word in a text is. However, raw frequency might be inadequate. For example, two documents of different length could not be compared adequately as the words in the longer one will probably have a higher frequency just because of its length. In other words, we could not say that a particular term t is more important in one of those documents compared to the other as their frequency is not comparable. To address these type of problems it is common to normalize the frequency of a document.

The tf-idf statistic, short for term frequency-inverse document frequency, is commonly used to normalize the terms on a text. The objective of this statistic is to measure the importance of a term. The higher its frequency on a document, the higher its value will be, but is offset by the frequency of the term in the set of the corpus. As its name suggests, it is composed of two statistics: tf (term frequency) and idf (inverse document frequency). The tf statistics can be computed directly as the number of times a word appears in the document. However, considering the previous example in which long documents will usually have a greater number of words, another function to compute term frequency is by normalizing frequency of the most occurring term:

$$tf(w, \mathrm{d}) = \frac{freq(w, \mathrm{d})}{max\{freq(w', \mathrm{d}) : w' \in \mathrm{d}\}}$$

On the other hand, the idf statistic measures how important the word is in the complete set of documents being analyzed. That is, it gives rare terms a higher importance than those that appear too frequently in the whole set of documents (D) to consider.

$$\mathrm{id}f(t, D) = log\frac{|D|}{1 + |\{\mathrm{d} \in D : t \in \mathrm{d}\}|}$$

Finally, the tf-idf is computed as:

$$tf - \mathrm{id}f(t, \mathrm{d}, D) = tf(t, \mathrm{d}) \cdot \mathrm{id}f(t, D)$$

Once the document is represented by a model, we can measure its similarity compared to other documents. As mentioned earlier, as the space model represents documents as vectors in space, the more similar they are, the closer they should be. The Euclidean distance is one of the metrics that can be used to measure the distance of two documents. It determines the dissimilarity of two documents: the greater the value of the metric, the farther they are in space. We compute this measure for documents a and b as:

$$\mathrm{d}(a, b) = \sqrt{\sum_{\mathrm{i},j=0}^{\|V\|}(a_\mathrm{i} - b_\mathrm{i})^2}$$

Another commonly used metric is the cosine, which looks at the angle formed between two vectors. It indicates the similarity between two documents and is defined as:

$$cos(a, b) = \frac{a \cdot b}{\|a\|\|b\|}$$

The main difference between both is that euclidean distance considers the length of the vectors and cosine focuses on their direction. However, if the vector representing a document is normalized, then both measures should be similar.

2.3 Geographical Data Related Concepts

Related to the goals of this thesis, it is relevant to define *spatio-temporal*, and *geopolitical*:

- **Spatio-temporal or geo-temporal**: regarding space, as a physical extent in all directions, and time.
- **Geopolitical**: related to the geographical and political factors that can influence to a location's power and relationships with other locations. Commonly, countries are considered as the administrative level for geopolitical analysis.

Regarding the process of extracting geographical data, the following concepts are pertinent to this dissertation:

- **Geoparsing**: the process of converting unstructured text containing name or description of places into unambiguous geographic identifiers such as coordinates or metadata.
- **Geotagging**: the process of adding geographical metadata to media, such as photograph or video. This term also refers to the process of obtaining geographical coordinates from non-coordinate geographical information such as address.
- **Toponym**: the name of a place or a name related to a name of a place.
- **Toponym extraction**: the process of extracting name of places from text, usually unstructured text.
- **Toponym resolution**: mapping an identified toponym to an structured representation of the geographical place that is referred to.

In particular, when dealing with big volumes of data, the processes of toponym extraction and resolution are aimed to be conducted automatically. For this goal there are several libraries available such as CLAVIN [24] or geodict [187].

Chapter 3

Literature Review

Our research on modeling and visualizing news events extracted from social media in their geo-temporal context corresponds to two research areas: Data Mining and Information Visualization. In this section we review relevant work in both these areas. Regarding the first one, we briefly discuss some relevant definitions of event and review research on event detection in text streams. Later, we describe in depth research on how social media has been used to analyze historical events. In addition, we discuss how this data has been modeled to achieve this type of analysis. Regarding the visualization aspect of this thesis, we first concisely describe related concepts about geovisualization. The following sections, detail research about general geotemporal visualization designs and on visual interfaces specifically designed to display social media information in their geotemporal context. Finally, we dedicate a section to glyphs and their use for displaying geotemporal data.

3.1 Data Mining related Work

3.1.1 Event Definition

The word "event" can be defined in different ways. Some generic definitions, like the one from the *Online Oxford Dictionaries*, define an event as "a thing that happens or takes place, especially one of importance" [138]. Others define it as something that happens or unfolds at a particular time and place [13, 199]. Regarding news events, Sayyadi et al. [160] define them as "*any event (something happening at a specific time and place) of interest to the (news) media*". In the context of social media, Dou et al [185] define "event" as:

> *An occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by topic and time, and often associated with entities such as people and location.*

In later work, they stated that "*Collectively, events serve as a succinct summary of social media streams. Individually, event and its sub-events, reveal the evolution of certain social*

phenomena over time" [53]. For the present thesis we consider that an event is observed given a disturbance in the social streams, reflecting that something happened at a particular time and place, and that was sensed and shared by the online social networks users.

3.1.2 Detecting Events from Social Media Data

Detecting topics in streams of traditional type of articles have been studied for years. In particular, in the Topic Detection and Tracking (TDT) project [12] researchers have studied the appearance of new topics in continuous news-streams, tracking their evolution and reappearance. At the beginning of the study, the project focused only on "topic" detection, but the concept was changed to "event", meaning something unique that happens at some point in time [12]. TDT focuses on two kinds of event detection: *retrospective event detection* and *on-line new event detection*. *Retrospective event detection* focuses on finding unrecognized events in a previously collected corpus of stories. This is done by grouping stories together that refer to a same event, assuming that each story describes one event at most. On the other hand, *on-line new event detection* is the task of identifying a new event from a continuous stream in real time. Each story is processed in sequence as they appear in the stream and the system decides whether it describes a new event or not. The system has all the stories processed before, but does not know the stories that will appear next [13].

Given the large amount of data generated from online social networks, much of the attention from the scientific community is now focused on their study. Working with *social data streams* represents an interesting challenge as it not only requires dealing with text processing but also with the network of users [7]. In particular, detecting events from Twitter gave rise to new problems since tweets are short in length, have an unstructured nature and are published by an heterogeneous group of people. These challenges have been addressed in the intensive research on event detection from social networks data, particularly from Twitter [18, 42, 83].

3.1.3 Quantitative Historical Event Analysis

We provide a revision of the literature on *quantitative history* research applied to event analysis and social media. Quantitative history is an approach to historical research that makes use of quantitative and digital tools [195]. To the best of our knowledge, our work is the first at the time of its writing (March 2018) to make use of social media data for quantitative historical research.

Prior work used digitized newspapers and books for extracting quantitative knowledge [129, 113, 35]. Michel et al. [129] built a corpus of 5 million books and analyzed them using word frequencies to investigate cultural trends, and called this type of study "Culturomics". Leetaru [113] performed a large-scale study of 30 years of digitized newspapers, described in the previous section. Chadefaux [35] used a dataset from Google News Archive to predict military conflicts.

A different line of research covers digitized writings and the Semantic Web. Suchanek and Preda [165] proposed the study of "Semantic Culturomics", in which the analysis of newspapers should go beyond the study of word frequencies in order to integrate knowledge bases (such as DBPedia [46]) to answer complex user queries. Additional research has used knowledge bases along with human writings, such as newspapers [88, 152]. Meroño-Peñuela et al. [126] provide a survey on this topic.

Compared to prior work, our approach is the first to consider user-generated information networks, such as online social networks, which are a growing data source at much larger scale. We consider that social media can provide additional and novel information to that found in news articles and books. User-generated content reflects social opinions and points of view related to current world-events. This content is generated in real-time, it is not edited and does not depend on the editorial lines of formal news outlets. We believe that these unique characteristics make social media a challenging and valuable source of historical information. Our approach incorporates the content of social media platforms about real-world news, as well as aggregated geographical information that conveys the importance and scope of these events.

3.1.4 Event Models Using Social Media

There is a lot of research in social media event analysis that has been directed towards the creation of event models for specific tasks such as summarization and characterization of events in social media streams.

For example, in the area of automatic text summarization, Chakrabarti and Punera [37], used hidden Markov models to represent sub-events, within a broader event that is described using Twitter data. This model identified sub-events based on the burstiness of the input data stream and the word distribution of the main event. Another approach was presented by Quezada and Poblete [150], which focused on automatic summarization of multimedia content by using social media posts as surrogate text for multimedia documents. A similar approach was used by Alonso et al. [14], which was based on the *social signature* of documents (that is, the set of keywords of social media messages that point to a document), to augment the document information.

The behavior of people surrounding an event is an important subject to study. Kalyanam et al. [91] studied how exogenous events, in this case real-world news, propagate in social media. In their work, they modeled news events based on the interarrival time between social media posts, without considering any of the geographical information associated to the event. Their goal was to model the intensity of the user activity that is triggered by a real-world news event.

The sentiment is also a studied aspect about human behavior of an event. In a different type of study, Leetaru [113] performed a large-scale analysis of 30 years of digitized news articles. The author computed sentiment scores and geolocation for each article. The study indicated that some critical events in the past, such as social revolutions, could have been forecasted by looking at sentiment scores over time. In addition, the author performed

community detection on country graphs by analyzing news in which two or more countries were involved. In this sense, our approach is similar, because we model countries in terms of their co-occurrence in news. However, our work focuses on automatic information extraction from online social streams and on the creation of a more general representation. We do not focus on the analysis of sentiment of edited content from formal news media outlets, but on the interactions between locations, based on the aggregated reactions and opinions of users of social platforms. Related to this work, Castillo et al. [33] study the related sentiment to the credibility of the information that is published in social media.

Finally, it is also important to mention that some research includes temporal features to, for example, detect events based on the temporal dynamics of their mentions in social media [78], and also for event categorization [151].

3.1.5 Geo-temporal Event Models Using Social Media Data

Several studies are available on the geographical and temporal characteristics of an event from social media. However, not much work has focused on high-level event modeling with context information of its spatio-temporal features. For example, in the work of Kamath et al. [92], Twitter *hashtags* (i.e., user-generated string prefixed by # that users add to tweets as a way to associate it with an event or a topic) were analyzed in a large-scale study of the spatio-temporal dynamics of *memes*. In this work a hashtag was represented as a tuple consisting of the coordinates of the hashtag's location over time. They used a simple model to find interesting insights about the adoption and spread of memes in social media. Memes are information which emerges from social networks and spreads in a viral way. However, meme dissemination does not necessarily resemble how other types of information will propagate, such as information about events that originate outside the network (i.e., exogenous events).

Certain studies focused specifically on the task of detecting events and tagging their relevant geolocations. In particular, some works targeted the detection of localized events [189, 3, 183, 110, 104], others, the detection of global events [157], and the detection of critical events [156, 47]. Dong et al. [50], specifically, considered that events had different temporal and spatial scales and proposed a multi-scale event detection approach for social media. This approach focuses on detecting and reporting events with geolocalization. Our current approach differs from existing work, in that we create an aggregated representation of the information about real-world events, producing a high-level representation that includes the event's geographical context, which is extracted from social media. In addition, we enrich the information about an event by using the locations of the users that post information about it.

Wang et al. [185] visualized topics based on the extraction of geographical entities from tweet text. They did not use this information to establish the location of an event, but rather for event exploration. SensePlace2 [121] is a Visual Analytics tool that allows users to explore a set of tweets and models them by showing two geographical types of information: the locations from where users discussed the topic and the locations being mentioned in tweets. However, unlike our work, this information was only used at single tweet level, and not at event level.

In the domain of cyber-physical systems, *events* are viewed as conditions of interest [168] within a cyber-physical system, or as the co-occurrence of two people in the same physical place [108]. In general, events are modeled according to the state of the objects in the system, considering attributes, time and location. The work presented by Tan et al. [168] bears certain similarities with our own, in the sense that they considered an event to encompass multiple information about a condition of interest in the system (in our case in the online social network), including time and physical locations. In addition, the authors defined different kinds of temporal and geographical scopes for their events, which are similar to our definition of *event impact*. The main difference is that our approach aims to capture high-level information of how a complex exogenous event, such as a news event, is perceived by social network users in an aggregated way. Therefore, we focus on geopolitical divisions as units of aggregated spatial information and on representing geopolitical interactions.

Despite that the idea of adding spatio-temporal context to social media data is not novel, to the best of our knowledge our work is the first that formally introduces *protagonist* and *interested* locations in a high-level event representation. The novelty of our approach relies on the extension of the notion of spatial context, first by associating real-world news to one or more protagonist locations, and second by associating real-world news to the locations where they generated interest. In addition, our work does not focus on event detection, classification or summarization, as most of the prior work on event analysis does.

3.2 Information Visualization Related Work

In this section we focus on literature related to this dissertation from the Information Visualization research area. We first describe some studies that focus on the theoretical aspects of geo-temporal data visualization. Later we review some general techniques to visualize geographical and geo-temporal data. We then proceed to describe some research that visualize geo-temporal data from social media. Finally, we review some work related to glyphs and how can they be used for displaying geographical data.

3.2.1 Geovisualization

Geovisualization is a research field that integrates approaches from visualization in scientific computing (ViSC), cartography, image analysis, information visualization, exploratory data analysis (EDA), as well as GIS [118]. It represents geographical data through visual tools for exploration, analysis, synthesis and presentation, for theories and methods development. With its base on cartography and map design, nowadays geovisualization is an area of activity that leverages geographic data resources to meet a very wide range of scientific and social needs.

There are several works that focus on the theoretical issues about geovisualization. For example, Andrienko et al. [15] reviewed existing techniques to extract properties of spatio-temporal data and the exploratory tasks they can potentially support. Regarding the first,

they classify spatio-temporal data in three main types:

- **Existential changes:** which refers to entities appearing or disappearing (e.g., news events)
- **Changes in spatial properties:** location, shape, size, orientation, etc (e.g., soil erosion)
- **Changes of thematic properties expressed through values of attributes:** qualitative changes and changes of ordinal or numeric characteristics (e.g., population increase o decrease).

In particular, they denote "events" as spatial objects undergoing existential changes, and distinguish them between momentary and durable events. Each of the previously described kind of data can be represented using different visualization techniques. For example, existential changes can be represented as nodes on a map and their appearance or disappearance when changing time values, like VisGets [51]; changes in spatial properties can be depicted with animation as weather forecasts; and changes of thematic properties can be illustrated by assigning different color shades to the inner shape of countries on a map to represent their population [118].

On the other hand, a visualization tool must enable users to conduct specific tasks. Regarding the tasks that can be conducted in spatio-temporal visualization tools, Adrienko et al. refers to the work of Peuquet [144] about the three basic questions a geo-temporal visualization can answer: *what?*, referring to objects to analyze; *when?*, referring to the time objects are valid, and *where?*, referring to the location or space where objects are. A user can conduct the task of finding one of them, by using the combination of the other two:

- *when* + *where* $\rightarrow$ *what*: Describe the objects or set of objects that are present at a given location or set of locations at a given time or set of times.
- *when* + *what* $\rightarrow$ *where*: Describe the location or set of locations occupied by a given object or set of objects at a given time or set of times.
- *where* + *what* $\rightarrow$ *when*: Describe the times or set of times that a given object or set of objects occupied a given location or set of locations.

Nowadays, several taxonomies can be found in the literature that describe more detailed tasks for geographical and temporal data. For example, Roth [155], empirically derived interaction primitives for geovisualization. In particular, he describes five objectives three broader user goals and three operand primitives: space-alone, attributes-in-space and space-in-time. On the other hand, Nusrat et al. [135] describes a task taxonomy focused specifically on cartograms.

3.2.2 Geographical and Geo-temporal Visualizations

Maps are the most commonly used visual representation of spatial data, which aim to reflect the real-world as much as possible. Symbols can be used to show additional information in maps, such as arrows to depict flow[145]. A map can be modified to represent other variables, such as Choropleths do, by representing data coloring or applying texture to regions. Cartograms are maps in which distances or areas are distorted in proportion to a variable of interest. There are several techniques to generate cartograms, which can be broadly categorized into four types [135]: contiguous, non-contiguous, Dorling [52], and rectangular [180]. Other abstract ways of displaying geographical data are bar charts, line charts and scatter plots[1], however these representations do not preserve geographical topology.

Including time in the analysis as a variable can increase the difficulty for designing an effective visualization. Several techniques are available in the literature depending on the task at hand. For instance, if the task is to compare how much time a car will take driving between two points in a city, there are some approaches that will distort the distance to reflect effective time used. Traffigram[87] is an example of this type of visualization, where a map is distorted to align its point to temporal equidistant contours from an origin.

A particular task involving geographical an temporal data is analyzing how geographical data varies over time. This is a common type of analysis performed on geo-temporal data. Roth [155] defined this task as an operand primitive called *space-in-time* and described it as interactions with the geographical component of the map, to understand how it changes over time. Two very popular techniques to do this kind of analysis are small multiples and animation. For example, in the work of Johnson et al. [90], several cartograms are displayed in sequence to depict the evolution of the Internet in the world between 1990 and 2013. On the other hand, animation is used to smooth the transition between views of geographical data in the work of Craig et al. [43]. Animation is not only used as transition between time frames, but also in combination of symbols to depict change. For example, in the work of Kim et al. [98], they include animated particles to create flow maps that are used to display geotemporal data without trajectories.

There are several sophisticated techniques that aim to show how geographical data changes over time. For example, the Space Time Cube [102] displays geographical information as a 2D map and displays change over time using the z-axis. The Space Time Cube has been used in several different areas such as analysis of earthquakes in the context of transient events [69], visualization of patterns of crime cluster [132], and analysis of dynamic lightning data [142]. Other works use pattern mining to extract relevant information, like Yusof et al. [203], where they used pattern mining algorithms to extract wind sequential patterns for visualizing them using a 3D wind rose and a space-time cube.

3.2.3 Geo-temporal Visualizations Using Social Media Data

There are several visualization tools that show where a news event has happened or from where social media users are commenting on it. In this section, we review the tools that are

relevant for our work, mainly focused on what type of geographical information they convey and what users can obtain from them.

If an event is represented as a set of documents, then one way of understanding this event is by using the documents metadata. There are several visualization tools based on this idea which show the geographical distribution of documents, allowing users to answer specific questions. Some examples are TwitInfo [123], Jasmine [189], and others [89, 149, 84, 44, 36]. Some systems provide filters for users to select documents published from particular places at particular times. For example, ScatterBlogs2 [27] is a visual analytics system for understanding messages from Twitter that allows users to interactively filter messages by their geographical and temporal context, using the coordinates from where the message was emitted. Also, Bosch et al. [57] created a system that aims to help users analyze social media using various sources, including search and filtering features for messages in their spatial and temporal dimensions. Other work, such as the one from Zhang et al. [205] visualize other characteristics of a topic or news event such as the sentiment analysis. By using an approach based on quantum mechanics, they propose an Electron Cloud Model (ECM) to visualize the evolution of sentiment analysis for news events from micro-blog data. Then they use map representation to observe the distribution of sentiment and retweets within Chinese regions. All of these systems use a map to display the geographical distribution of messages, (or of users) in order to describe a topic or event. al. [103]. Whisper [32] uses a different metaphor: by representing messages of an event as seeds of a sunflower, a user can follow how information disseminates by viewing the locations from where people commented on an event, or from where a message was re-posted. In contrast to these approaches, which are centered on user messages, our visual tool focuses on the characteristics of an event as a whole, providing details (messages) on demand.

There are also visualization systems for describing events. Visgets [51] provides a visual interface to represent entities from different data sources, such as the ACM WWW proceedings or the social news site Global Voices Online [71]. A user can search and filter entities by time, space and keywords. Visgets represents entities by their geographical location using entity metadata. LeadLine [185] is an interactive visual analytics system that supports the exploration of events detected automatically from news and social media. The LeadLine system extracts places mentioned in messages to identify where a piece of news was relevant. Event Registry [58] is a system that monitors media sources to detect news events in more than ten languages. It also presents a map visualization that displays each event as a bubble over the location where it happened. In the work of Kraft et al. [103], they use a 3D representation of the world to display the amount of times some locations (e.g. cities) are mentioned in tweets in order to understand the "*where*" of a news event. Our approach complements these systems by leveraging event information and the impact that its information had in social media.

SensePlace2 [121] is a web system that shows locations mentioned in tweets and the locations from where these tweets were published. From that point of view, SensePlace2 is a system that allows users to ask: *"what places are involved in an event and from where are people commenting on it?"*. Therefore, it is the most similar system to the work presented in this paper. However, as the authors of SensePlace2 described in their own work, the main limitation of their tool is that it focuses more on the dimensions of the events rather than

on the events themselves. Our work complements SensePlace2 by: 1) focusing on overall event information, 2) by allowing users to explore relationships among countries, and 3) by showing the user if news events are local or international. Another system that is similar to ours is The News Co-occurrence Globe [70], which displays the co-occurrence of countries in news media reports on a 3D map. However, it does not currently provide the functionality to put focus on events. Our work allows the user to focus on events to see how relationships between countries are created and evolve over time.

In summary, to represent events in their geo-temporal context, most visualization systems either show the geographical distribution of the documents that discuss news events, or the information about the event itself. However, these approaches are limited if the user needs to retrieve news events or ask complex questions such as *where did event x happen?*, *how did people around the world react to event x in social media?*, *did event x impact only locally or did it have global impact?*, *which countries showed the most interest in event x?* or *have other countries also been involved in similar events to x?*. In particular, to the best of our knowledge, our is the first approach to consider that events can be related to multiple locations, reflecting interactions between geopolitical entities. Overall, our tool is the first to allow historical news exploration and retrieval that considers the temporal and spatial context of the user and of the event. In addition, providing the means for manual exploration of vast amounts of contextualized events described using social media data.

3.2.4 Glyphs for Multivariate and Geographical Data

A glyph can be broadly defined as a visual representation of a data unit, for which its graphical attributes have been mapped from the attributes of the data record [186]. More narrow definitions describe them as small independent visual objects that depict a data record. These objects are positioned in space, as visual signs that differ from other types of signs such as icons or symbols [26]. Glyphs are commonly used to represent multivariate data, in which each attribute of the data is mapped to a visual variable of the glyph such as shape, color or size. By displaying several glyphs, each of them representing an individual data record and grouping those with similar characteristics, a user can observe patterns in the data.

Using glyphs for representing multivariate data can have several benefits. Bürger and Hauser[63] describe the ability of glyphs to incorporate multiple dimensions at once. In addition, they discuss that it is possible to combine them with other visualizations. However, as stated by Borgo et al.[26], some studies indicate that only well designed glyphs are actually useful. Indeed, several guidelines or considerations have been proposed in order to provide principles that lead to effective glyph designs. For example, Chung et al. [40] describe eight design principles for the creation of sortable glyphs: typedness, visual orderability, channel capacity, separability, searchability, learnability, attention balance, focus and context. Karve and Gleicher [95] identified three considerations for designers of complex and compound glyphs: integral-separable dimension pairs, referring to the integration of multivariate visual variables; natural mappings, which denotes the need of a clear relationship between the visual attributes and the data attributes; and the perceptual efficiency of the encoding, which regards with the most efficient choice of a visual variable for a particular type of attribute.

Lie et al. [115] propose guidelines for glyph-based 3D data visualization.

There are several generic glyph designs such as star glyphs, Chernoff faces [38] or the autoglyphs [21]. On the other hand, more specific designs have been used for applications in medicine [137, 128], network security [140], sports [114], among others.

Particularly regarding geographical data, glyphs are commonly used as symbols over maps to display one or more variables. For example, Lucchesi et al. [120] used them to display uncertainty of poverty level, Sanyal et al. [158] for displaying meteorological uncertainty in combination with spaghetti plots, and Villanueva et al. [182] for representing stream data for Smart Cities (like pollution). On the other hand, other techniques combine arrow glyphs with color to represent magnitude and direction of a vector field [29]. However, despite being a popular way to display multivariate data, they have not been commonly used as the visual representation of a geographical location. In this thesis we present a new way of using glyphs for geographical data, which represent the world as a reduced visual element for easier comparison of several time frames.

Chapter 4

Geo-temporal Representation of Events Extracted from Social Media[1]

Twitter provides excellent conditions for social behavior analysis, as well as comparative historical research, among many other social and scientific disciplines. In particular, the field of *comparative historical research* examines historical events in comparison to other historical events to gain general knowledge that goes beyond a particular event [190]. So far, historical research had been restricted to traditional archival data and historians' written account of past events. Despite the usefulness of historical information extracted from social media, there is not much research addressing the topic of retrospective analysis of this data. Social media in general, and Twitter in particular, produce huge volumes of streaming data that is volatile, which most likely explains why existing research concentrates mainly on event detection, either targeting events in general or focusing in particular type of vents such as natural disasters.

To represent events from data extracted from social media, several work in the literature describe them by their more representative words [143]. Additionally, some of them include the date when they were detected, or even the date an event is expected to happen [151]. Some include Twitter-specific features such as hashtags, retweets or followers [147, 20]. When considering the geographical characteristics of an event, most event representations either focus on the geographical propagation of messages discussing it [92], or the location where an event happened [157, 112, 184, 4], in which case only one is usually considered. These representations are normally directed to understanding particular events more than the political and historical implications of them.

To include the political aspect of the geographical features of an event, we present a novel high-level event representation, called *spatio-temporal context-aware event representation*. The main purpose of this representation is to provide means to gain insight about real-world news from social media as well as from the relations between locations and impact that these events induce. Specifically, we formally define our event representation and describe how it can be used to study relations among locations.

[1]The contributions of this chapter were made in collaboration with PhD Student Mauricio Quezada Veas and was published in the work of Peña-Araya et al. [141].

We consider that the physical location where an event happened in the real world and the propagation of messages discussing it in social media are pieces of information that are incomplete in isolation. For this reason our event representation incorporates these two types of spatial data about an event: 1) locations directly *involved* in the real-world event occurrence (i.e., the main places that are mentioned in messages about the event), which we refer to as *protagonist locations*, and 2) locations from where social network users *comment* on the event (i.e, the places where users that comment are located), which we refer to as *interested locations*.

Let us consider the earthquake that took place in Nepal in April, 2015 [193]. For this event, most of the messages mentioned Nepal, which indicated that this was the location where the event had taken place. Therefore, if we consider locations at a country level, Nepal can be regarded as the *protagonist location* of that event. However, the users that posted the messages about Nepal were distributed all over the world, indicating that this event had global impact. Furthermore, some countries had more users interested in the event than other countries, such as, neighboring countries and countries with citizens among the victims. These would be considered as the *interested locations* of that event. It is important to note that the *protagonist locations* not necessarily involve the physical locations where an event happened. For instance, let us consider the soccer match in the FIFA World CUP of 2014 between Costa Rica and Greece on June 30th, 2014 [64]. For this event, most messages mainly mentioned both countries so they can be considered as the *protagonist locations*, while Brazil, the physical country where the match took place, was not mentioned enough to be considered a protagonist.

Our work is based on the hypothesis that by adding geo-temporal context to news events, such as protagonist and interested locations, and the time at which it occurred, we can discover new information based solely on social media data. In particular, the application of our event representation allows us to find relationships among events and among locations, such as:

(i) **event similarity:**

- **based on their protagonist locations**, i.e., retrieve all the events that occurred in certain location, or that directly involved similar groups of locations;
- **based on the locations that are interested in the event**, i.e., retrieve all of the events that produced the similar interest in other locations.

(ii) **location similarity:**

- **based on events in which a location is protagonist**, i.e., retrieve the locations that are protagonists in the same events;
- **based on their interest in events**, i.e., retrieve sets of locations that showed similar levels of interest in the same events.

(iii) **any combination of the above.**

These similarity relationships along with temporal context can facilitate the implemen-

tation of novel information retrieval tasks. These tasks include: event search, event understanding, geopolitical analysis, international relations analysis (when considering locations at a country level), historical comparative analysis, among others.

4.1 Event Representation Definition

We represent an event as a complex information unit that encompasses all of the available social media content associated with a certain news topic, as well as its aggregated spatial and geographical information.

In particular, we incorporate information about the locations involved in the event occurrence, and the locations of the users that post messages about the event. This representation is solely based on the social media activity surrounding the event in the online social network, without including any external information sources.

Specifically, we define two types of spatial contexts, which we call:

1. **protagonist locations**, which are the locations involved in the event, and

2. **interested locations**, which are the locations from where users comment on the event.

For example, let's consider the news about Chile and Peru's maritime dispute at The Hague in The Netherlands [19]. If we define locations at country level, then this is an event for which Twitter users mention mostly three countries when discussing the event: Chile, Peru and The Netherlands (other mentions are negligible). Hence, according to our definition this event is considered to have three protagonist locations. However, users that comment on this event are located mostly in: Chile, Peru, Argentina and Bolivia. Therefore, the event is considered to have four interested locations.

More formally, we define an event E as a tuple of the form:

$$E = (K, D, T, \mathbf{P}, \mathbf{I}) \tag{4.1}$$

where K is a set of keywords, which succinctly describe the news topic, D is the date of the event detection, T is a set of tweets about the event, published by users of online social networks. In addition, consider $L = \{l_1, l_2, \ldots, l_{|L|}\}$ to be the set of existing locations. We augment the information about the event by explicitly including its spatial context with the vectors $\mathbf{P}$ and $\mathbf{I}$, which correspond to the *protagonist* and *interested* location values, respectively, for the event E. This is, the j-th dimension of vector $\mathbf{P}$ contains the number of times that the location l_j is mentioned by the tweets in T. On the other hand, the j-th dimension of vector $\mathbf{I}$ contains the number of tweets in T that were posted by users in the location l_j.

As our event representation should allows us to analyze the political aspects of the geographical features of an event, we consider locations as categorical entities that have a geographical reference. In other words, we consider that a location can be of any type of geopolitical division granularity, such as a city, a region, a country, a continent, etc, instead

of the coordinates of a point or area. For example, if we focus on the country administrative level, we will consider that set L is composed by all the countries of the world. On the other hand, if we work on the regional administrative level of Chile [196], the possible protagonist or participants locations would be among the 15 regions of that country.

Using the information introduced by vectors $\mathbf{P}$ and $\mathbf{I}$ we can derive the scope of an event E from two perspectives, *provenance* and *impact*, defined as follows:

- **Provenance:** Indicates whether an event is local, regional or global in terms of the locations it involves. We consider an event to be of *local provenance* if it involves only one protagonist location. We consider an event to be of *regional provenance* if it involves two or more protagonist locations that are all from a same region (e.g., for countries, this means neighboring countries or from a same continent[2]). We consider an event to be of *global provenance* if it involves two or more protagonist locations in which at least one is not from the same region. Vector $\mathbf{P}$ contains this information for a given event E.

- **Impact:** Similar to provenance, this vector indicates if an event is local, regional or global in terms of how many locations show interest in it. We consider an event to be of *local impact* if it generates conversation from users in only one location (i.e., one interested location). We consider an event to be of *regional impact* if it generates conversation from users in more than one interested location, all from the same region. We consider an event to be of *global impact* if it generates conversation from users in more than one interested location in which at least one of those locations is not from the same region. Vector $\mathbf{I}$ contains this information for a given event E.

For example, the message *"Australia confirms signals detected by China 'consistent' w/ #MH370 black box"*, discussed an event in which Australia and China are involved. Therefore, Australia and China can be considered as protagonist locations in this event. On the other hand, this particular news event was discussed extensively by users located in several countries, including: USA, Canada, Colombia, U.K., India, Nigeria, South Africa, Indonesia, Australia, France, Germany, China and Italy. Therefore, this is an event that had *global provenance* (i.e., more than one protagonist location from different regions) and *global impact* (i.e., more than one interested country from different regions). It should be noted that there can be different levels of "global impact", depending on how many different locations show interest in the event (e.g., high-impact global world events will spark conversation in many countries).

For the context of this thesis, we work with locations at the country administrative level. Therefore, at times we use the concepts of "locations" and "countries" interchangeably. In particular, in the following section, we define a representation for relations among locations, which we exploit to gain data insight about international relations, as described in following chapters.

[2]According to the division that considers 7 continents: Asia, Africa, Europe, North America, South America, Antarctica and Australia.

Representing Relations Among Locations

The spatio-temporal context-aware event representation allows us to extract different types of relationships among locations for a given event collection. In particular, we define a *protagonist-interest* vector **pi** for a location l_j, which represents the interest that other locations have in events that have l_j as a protagonist. We define **pi** for l_j as:

$$\mathbf{pi}(l_j) = \Big[w(l_j, l_1), w(l_j, l_2), \ldots, w(l_j, l_{|L|})\Big] \tag{4.2}$$

where,

$$w(l_j, l_k) = f(\#\ of\ events\ that\ have\ l_j\ as\ protagonist\ in\ which\ l_k\ shows\ interest),\ \forall\ l_j, l_k \in L \tag{4.3}$$

Likewise, we also define the *co-protagonist* vector **cp** for the location l_j as follows:

$$\mathbf{cp}(l_j) = \Big[w'(l_j, l_1), w'(l_j, l_2), \ldots, w'(l_j, l_{|L|})\Big] \tag{4.4}$$

where,

$$w'(l_j, l_k) = f(\#\ of\ events\ l_j\ as\ protagonist\ in\ which\ l_k\ is\ also\ a\ protagonist),\ \forall\ l_j, l_k \in L \tag{4.5}$$

The relationships between locations, given by **pi** and **cp**, allow us to identify similarity relationships among locations, such as:

- **Locations that produce similar interest:** from **pi** we can extract sets of locations (countries) that are similar, based on the level of interest that they produce in other locations (countries). For example, they can be obtained using k nearest neighbors or by clustering locations' **pi** vectors.
- **Locations that are protagonists of the same events:** from **cp** we can identify which locations (countries) are similar, based on their interactions (i.e., they are protagonists of the same event) with other locations (countries). For example, they can be obtained using k nearest neighbors or by clustering locations' **cp** vectors.

The weights, $w(l_j, l_k)$ and $w'(l_j, l_k)$, are expressed as a function $f(x_{j,k})$, where $x_{j,k}$ corresponds to *# of events in which l_j and l_k interact*. In particular, for the visual tool described in Chapter 5, users have expressed the preference of visualizing the *absolute* number of events in which two countries interact (i.e., $f(x_{j,k}) = x_{j,k}$). Nevertheless, there are other cases in which the analyst could prefer the weights to reflect the fraction of events in which two countries interact in relation to the total of events for one of the two locations (e.g., $f(x_{j,k}) = x_{j,k} / \max(\#\ of\ events\ in\ which\ l_j\ or\ l_k\ participate)$). This can be useful in cases that the number of events in which different locations participate are very concentrated on specific locations. We explore cases such as these in following chapters that describe the data insight we extracted using the model.

We note that weights can also be expressed as functions of the *# of tweets* or the *# of users*, and in addition, the proposed representation allows us to also specify *interest-interest* and *interest-protagonist* vectors, in a similar fashion to **pi** and **cp**. However, we do not focus on these variations at this moment.

4.2 Exploratory Analysis Using the Event Representation

We present an exploratory data mining analysis that uses the information provided by our spatio-temporal context-aware event representation. We describe our empirical findings, which illustrate the usefulness of our proposed event representation. This analysis considers the location context of events at the country-level geopolitical division. This allows us to explore the international interactions given by our current dataset. One of the first steps to conduct this analysis was to consider that some countries could be more represented than others in our dataset which is explained in the next section. Later, we describe the geographical coverage of events and the international relationships that were generated as consequence of these events.

4.2.1 Empirical Setup

We provide an overview of the data extraction methodology that allowed us to generate the dataset used for the exploratory analysis presented in this chapter. This setup is divided in two main components: the news events extraction and the geographical context extraction. Both are responsible for the creation of the input dataset from which the event representation is created in the following step. Given that event detection and extraction are beyond the scope of this thesis, we chose to rely on an existing approach that retrieves a set of events that are comprehensive and cohesive enough to test our event representation. Nevertheless, we acknowledge limitations in the type of events collected by this setup, discussed in Section 4.3, but we believe that these limitations do not affect the generalization of the results of the proposed event representation.

News event extraction setup. The news event extraction module corresponds to that used by Kalyanam et al. [91], which consists of an ongoing process that periodically retrieves tweets about real-world news. We provide an overview of this process, which produces coherent sets of tweets about news topics, although with certain degree of noise that is well tolerated by our system. In particular, this is a two-stage iterative process that consists of 1) *news topic identification* (i.e., detection), and 2) *event tweet extraction*. We describe them briefly next (more details on this method, including the validation of the cohesiveness of the resulting events can be found in Kalyanam et al. [91]):

1. **Topic identification.** This approach does not detect events directly, but rather restricts itself to topics that have been posted on Twitter by mainstream news media accounts. The system periodically (each hour) retrieves headlines posted on Twitter

by a set of *seed* news accounts, which must be provided. Using association rule analysis over the set of headlines collected in the cycle, the system outputs high-support sets of keywords ($\{K_1, K_2, \ldots, K_n\}$). These sets of keywords constitute terms that were posted together in a headline by more than one news outlet within an hour.

In this particular setup, the seed set of news accounts correspond to 55 well-known international news media outlets (with verified accounts). These accounts are mostly from English-speaking sources based in the United States and Great Britain, such as @BreakingNews, @CNN, @NYTimes, @Jerusalem_Post, @AJEnglish, @NDTV, etc.[3]

2. **Data collection.** This stage iteratively takes the keyword-sets produced in (1), and uses each keyword-set $K \in \{K_1, K_2, \ldots, K_n\}$ to query the Twitter Search API in order to retrieve tweets T from *regular users* that also contain the keyword-set (i.e., that are commenting on the same news topic as the headlines). The search is done within the same hour in which the headlines were retrieved, removing tweets that where more than a few hours old, narrowing down the number of tweets that do not belong to the news topic due to the temporal relevance of the event. In principle, each keyword-set K is considered to be related to a unique news topic E. However, several keyword-sets could be referring to a same news topic (within one cycle or across several collection cycles), therefore an additional step is applied to merge one or more set of tweets into one within a one-day time window.

Geographical context extraction setup. We create a methodology for extracting the protagonist and interested locations, as well as their frequency for an event E with a set of tweets T.

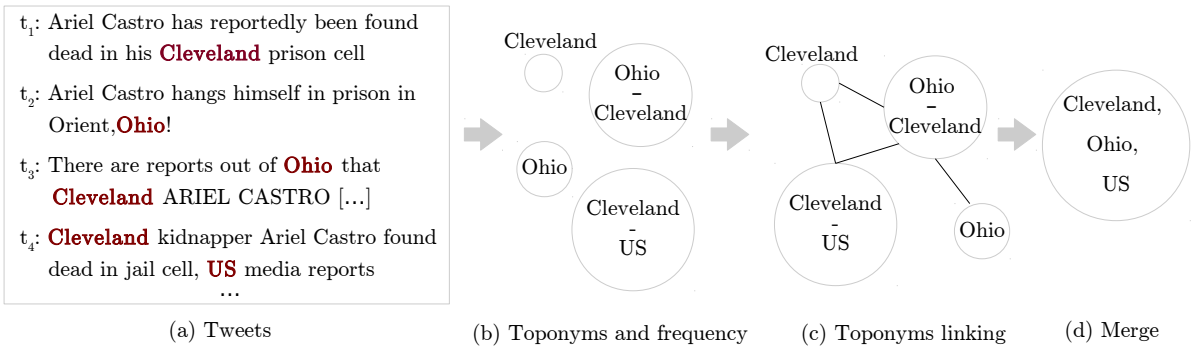
Recognizing the protagonists locations is directed towards understanding semantically where an event happened and/or which countries were involved in that event. To extract the location of an event, most of the existing research rely on two main sources: document metadata [159, 156, 3] or the document content [99]. One of the main problems with using the metadata of documents to geolocate an event is that the distribution of users who produce them can be biased. Therefore, events that occur in locations with less adoption of Twitter could be incorrectly geolocated. In particular, we believe that for general news events, user locations can be more useful for descriptive purposes than for geolocation extraction. On the other hand, when considering the content of documents to extract locations, it is hard to disambiguate toponyms (location names) within short text as they do not contain enough contextual information. Indeed, although there are several geotagers that extract locations from text [74, 188, 94], additional techniques must be applied to resolved them like, for example, using external information from sources like Wikipedia or ontologies [31].

In order to extract the protagonist locations reliably from social media data, which is by nature noisy, we need to analyze the complete set T of tweets for an event E. Let's consider the messages from Figure 4.1.a. In this case the word "Cleveland" can be incorrectly resolved as the area in northeast of England when analyzed in isolation. If this happens, the event could be incorrectly modeled as an international news as some tweets will refer to England

[3]Each Twitter account can be accessed in `https://twitter.com/accountname`, where `accountname` is the name of the account.

and some tweets will refer to the United States.

To overcome this problem, we implemented a methodology that resolves toponyms by extracting its contextual information from within the message itself, as well as the text contained in all of the tweets in T in E. The methodology is summarized in Figure 4.1. We start by extracting toponyms from each tweet t_i, grouping toponyms that co-occur in the same message, and computing their frequency of appearance. For example, in Figure 4.1a, t_4 contains both "Cleveland" and "US", therefore we group them together, shown in Figure 4.1b. Then, the methodology links groups of toponyms that share a location name, shown in Figure 4.1c. Then, each toponym that resolves to a valid geographical location is merged into one toponym, (Figure 4.1d), and this final toponym is resolved to a location l. Next, location l is assigned to $\mathbf{P}$ in E if its frequency is greater than the defined threshold k (see Section 5.2).



(a) Tweets (b) Toponyms and frequency (c) Toponyms linking (d) Merge

Figure 4.1: Event geotagging methodology. (a) Given the set of tweets containing location names, (b) it first recognize them and extract the frequency of their appearance, (c) it later recognize those locations that have toponyms in common and link them. (d) Finally it merge those that correspond to the same location.

Both the toponym extraction and resolution phases are carried out using the off-the-shelf geoparser, CLAVIN [24]. This geoparser identifies location names in unstructured text and resolves them against a gazetteer to produce data-rich geographic entities. The gazetteer is downloaded from GeoNames [72].

Empirically, we observed that the precision and recall of the locations considered to be protagonist of an event depends mostly on a ratio, which we call α. For an event E that contained more than one location, we defined α as the minimum percentage of tweets that must refer to a location l_i in relation to the most mentioned location l_{max}, in order for l_i to be included in $\mathbf{P}$ or $\mathbf{I}$ vectors. Figure 4.2 shows an empirical analysis of the effect of α on the precision, recall and F1 metrics of protagonist locations on a sample of 100 events. Precision and recall were estimated based on a manual assessment of the protagonist locations of those events. Based on this variation α can be set as the value that provides the best tradeoff between F1 and recall ($\alpha = 19\%$ in our experiment).

In order to obtain the set of *interested locations* of an event E, we access the GPS coordinates of each tweet in T, or if this not available, the *location* of the user's profiles information associated to the tweet. The location of the user can be either GPS coordinates, set by

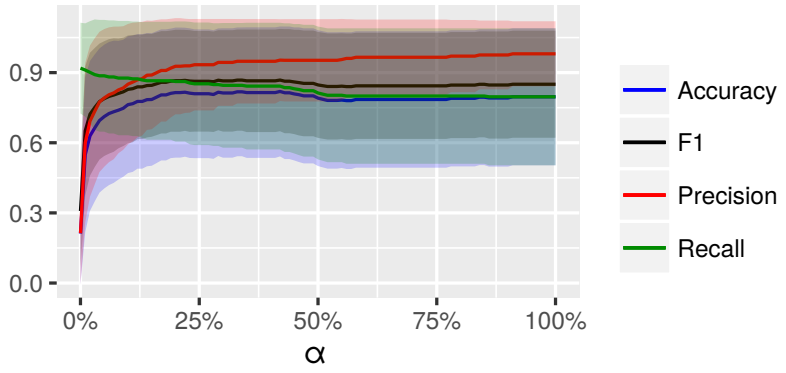


Figure 4.2: Mean and standard deviation of multi-label scores of accuracy, precision, recall and F1 measure by α ratio for 100 randomly selected events from our dataset.

the user's mobile phone, or natural text. (e.g., "Santiago, Chile"). If GPS coordinates are available the corresponding location is resolved in using the geodict library [187]. Otherwise, if the user has provided natural text as its location information we resolve it using CLAVIN.

Dataset Description. Using the previously described data extraction techniques, we collected a dataset of news events spanning from August 2013 to June 2015. This dataset consisted of 20,066 news events, which contained 193,445,734 tweets produced by 26,127,624 different users.

We note that our event representation and applications are independent of the data extraction methodology. Therefore, in order to improve the representativeness of our event collection in the future, less biased methods of event extraction can be used, such as automatic event detection techniques [127, 39] and/or the integration of more comprehensive sets of seed news sources, as done for Chilean news analysis by Maldonado et al. [122].

4.2.2 Country Representation Bias

This process was based on a seed set of internationally renowned news media accounts that publish information in English. This introduced a certain bias in our event collection towards events that took place in English speaking countries, and towards including more tweets in English than in other languages. For example, for the event *"correspondents dinner"* our current method will mostly retrieve tweets in English from users world-wide. On the other hand, an event described with a set of keywords which includes *"Barack Obama"* will retrieve tweets in several languages.

These biases must be taken into consideration because they can limit the representativeness of the findings yielded by our data mining analysis. Nevertheless, we believe that they do not invalidate our results, which show the perspective of a subset of the social network that is centered on news reported in the United States and Great Britain. Therefore, our results reflect the world-view of these two overly represented countries in particular, and of

(a) The United States is a protagonist country of the majority of the events, followed by Great Britain, India, Ukraine, Russia, Australia, Syria and Iraq. A darker color corresponds to a higher level of protagonism.

(b) The United States is the country that showed the highest level of interest in other events, followed by Great Britain, India, Canada, Nigeria, Indonesia, Australia, and South Africa. A darker color corresponds to a higher level of interest.

Figure 4.3: Summary maps of interest and protagonists.

English-speaking users in general. Furthermore, other studies using the full Twitter stream, such as that of Poblete et al. [148], show a similar data distribution to ours, indicating that this type of bias could be inherent to Twitter itself.

Furthermore, an in-depth exploration of the bias in our dataset showed that the number of tweets produced during an event did not depend on the number of seed accounts that covered that event. Our analysis showed that only 13.5% of the users in the entire collection had actually reposted a tweet from the seed news media accounts, which gives the overall impression that these accounts did not influence much the amount of interest expressed by users. Also, we found no relation between the number seed accounts that shared an event and the number of countries that participated in the event in terms of provenance or of impact.

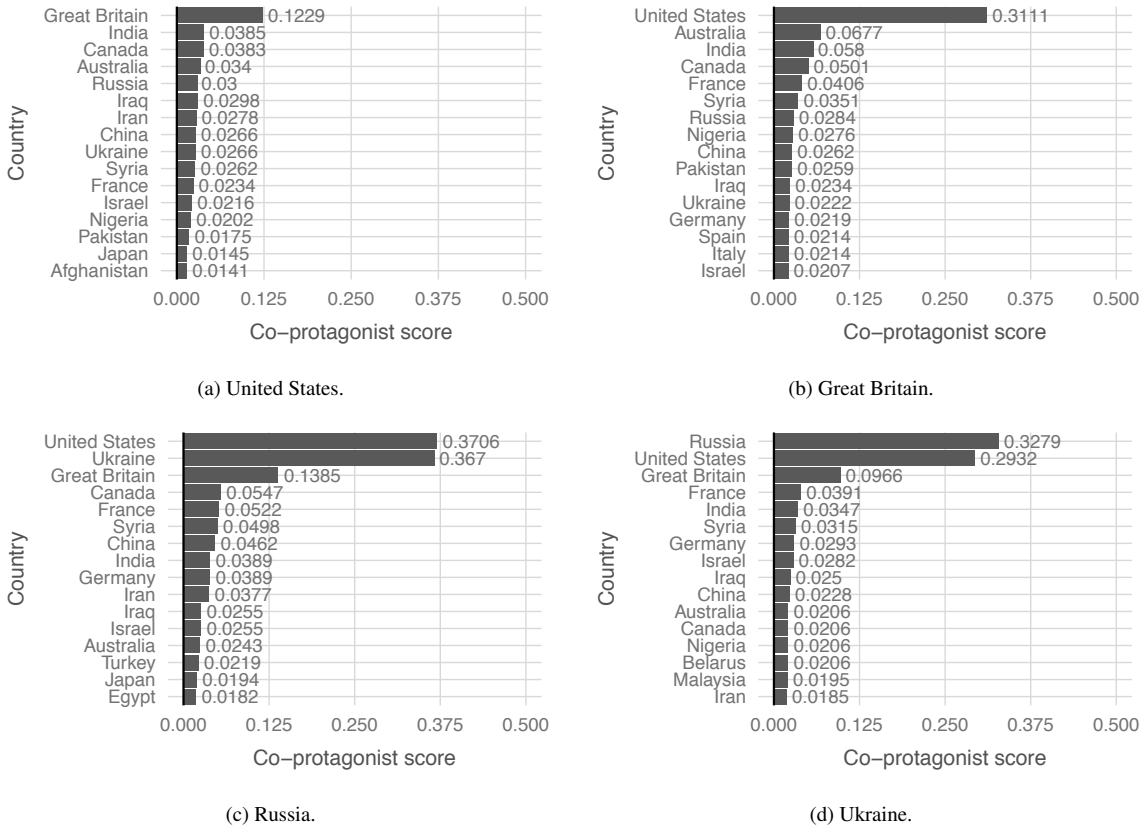
As mentioned in Section 4.1 we used a normalization for vectors **pi** and **cp**, defined in Equations 4.2, 4.3 and 4.4, 4.5, respectively. This normalization allows us to compare protagonist-interest and co-protagonist vectors in a way that mitigates the bias of overrepresented countries. In particular, for the **pi** vector we defined $w(l_j, l_k)$ as:

$$w(l_j, l_k) = f(x_{j,k}) = \frac{x_{j,k} - \mu(\mathbf{x}_{\cdot,k})}{\sigma(\mathbf{x}_{\cdot,k})}$$

and for **cp** we defined $w'(l_j, l_k)$ as:

$$w'(l_j, l_k) = f(x'_{j,k}) = \frac{x'_{j,k}}{x'_j},$$

where $x_{j,k}$ was the number of events that have l_j as protagonist in which l_k is interested; $\mathbf{x}_{\cdot,k}$ is the vector containing the number of events in which location l_k is interested, $\forall l_j \in L$; μ and σ are the mean and standard deviation of the distribution of events, respectively; $x'_{j,k}$ is the number of events for which both l_j and l_k were protagonists, and x'_j is the number of events that had l_j as protagonist.



(a) United States.

(b) Great Britain.

(c) Russia.

(d) Ukraine.

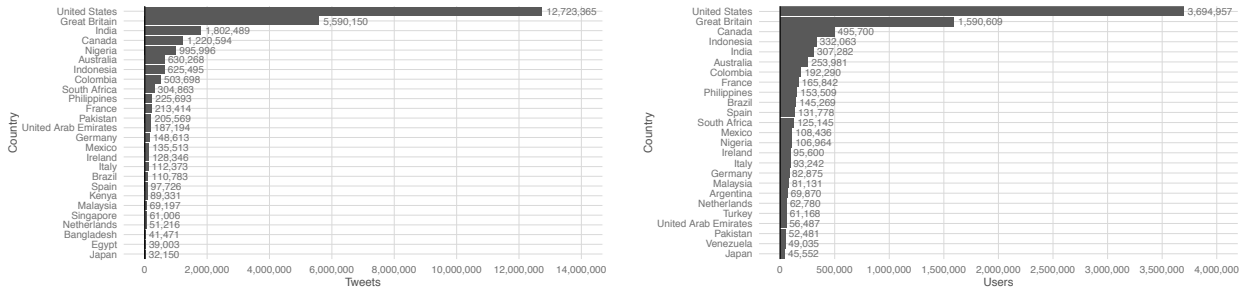
Figure 4.4: Relative co-protagonist measure of selected countries.

4.2.3 Geographical Coverage

We started by characterizing the spatial distribution of our collection to describe its representativeness in terms of geographical coverage. In terms of protagonist locations, the United States and Great Britain were the protagonists of the majority of the events, followed by India, Australia, Ukraine and Russia (Figure 4.3.a). The median number of events in which countries were protagonist is 18.5, indicating that only a few countries were the protagonists of the majority of events. Figure 4.3.a shows the distribution of the number of events in which countries were protagonists. When we computed the $\mathbf{cp}(c_{\mathrm{i}})$ vectors for selected c_{i} countries (Equation 4.4, normalized by the number of events in which a country c_{i} is protagonist), we observed that the United States and Great Britain were the protagonists of the majority of international events (Figure 4.4). There are some exceptions, such as Ukraine, which had only Russia as the co-protagonist of many of its international events (Figure 4.4.d).

In terms of worldwide interest, the countries that displayed interest in most events were the United States, Great Britain and India (Figure 4.3.b). In addition, these countries also contributed the most tweets (Figure 4.5.a).

We determined the location for 37.3% of the users (9,738,538 out of 26,127,625 users). These users were mostly distributed among the United States and Great Britain, followed by Canada, Indonesia and India (Figure 4.5.b).



(a) Tweets published per country.

(b) Unique users per country.

Figure 4.5: Description of the bias in the number of tweets and users, per country.

4.2.4 International Relations Exploration.

We explored the dataset in order to identify similarity between countries according to the events in which they are co-protagonists and the interest shown towards these events by the rest of the countries in the world. We found that applying standard similarity metrics over the data in the event representations, yielded relationships between certain countries that resemble intense historical interactions and/or geographical proximity.

In terms of protagonist locations, we found countries that were *similar*, meaning that they were protagonists of the same events. In this case we used the Jaccard similarity between each pair of countries as our similarity measure, representing each country by the set of events in which it was a protagonist. The Jaccard similarity between two sets x and y is defined as $sim_{x,y} = \frac{|x \cap y|}{|x \cup y|}$. We filtered out the countries that were protagonists of less than 130 events (corresponding to the 80-th percentile of events for which countries were protagonist).

We studied the distribution of our similarity metric, in order to determine which relationships between countries were significant. We fitted the *similarity* to a theoretical probability distribution using the R package *fitdistrplus* [4] and we found that the best fit was a Gamma distribution with parameters $shape = 0.8721$ and $rate = 85.7683$. Based on this analysis, if S is a random variable with a Gamma distribution representing the similarities between countries, then we defined the similarity between two countries x and y as being *significant* if its value was in the 95-th percentile of the distribution, (i.e., if $P(S < sim_{x,y}) > 0.95$). Using this criteria, we determined a similarity threshold of $sim^* = 0.032$, above which we considered its value to be significant. This threshold can be parameterized at the 80-th, 90-th, or 99-th percentile, as the researcher finds appropriate. Table 4.1 shows the top-20 most similar countries based on this similarity, making it to the 97.181 percentile of our dataset.

We found that Israel and Palestine were the most similar countries, followed by Russia and Ukraine, North Korea and South Korea, Great Britain and the United States, and Iraq and Syria (Table 4.1). Their similarities are higher than the 99.25% of the pair-wise similarities in our dataset. There are real-world historical and geographical relations between those countries that can account for these similarities (for example, the Ukrainian crisis [192], or the Israeli-Palestinian conflict [191]). On the other hand, some of the similarities can

[4] `https://CRAN.R-project.org/package=fitdistrplus`

Country i	Country j	x'_{i}	x'_j	Similarity	Percentile
Israel	Palestine	561	360	0.2863	99.969
Russia	Ukraine	823	921	0.2094	99.906
North Korea	South Korea	158	179	0.1866	99.843
Great Britain	United States	4,015	10,162	0.0966	99.248
Iraq	Syria	654	647	0.0833	99.092
India	Pakistan	1,561	453	0.0753	98.998
Iran	Israel	496	561	0.0698	98.841
China	Japan	646	354	0.0605	98.340
France	Germany	627	371	0.0583	98.184
Argentina	Brazil	130	236	0.0578	98.152
Australia	Great Britain	974	4,015	0.0577	98.090
Brazil	Germany	236	371	0.0575	98.058
Syria	Turkey	647	198	0.0536	97.964
Iran	Iraq	496	654	0.0512	97.777
Australia	Malaysia	974	262	0.0492	97.682
Argentina	Germany	130	371	0.0481	97.620
Australia	India	974	1,561	0.0475	97.495
Germany	Greece	371	155	0.0457	97.401
Canada	Great Britain	715	4,015	0.0444	97.275
Egypt	Libya	316	253	0.0440	97.213
Great Britain	India	4,015	1,561	0.0436	97.181

Table 4.1: Most similar countries in terms of being protagonists of the same events (***co-protagonist*** vector), using Jaccard Similarity. x'_{i} is the number of events in which country i was a protagonist.

be explained by the preponderance of certain events, such as the 2014 FIFA World Cup. These results indicate that there is information in Twitter data about real-world geo-political interactions which can be further studied using our event representation.

In Figure 4.6, we present three graphs where countries represent nodes and edges are weighted based on the Jaccard similarity. As we increase the threshold to connect two countries with an edge, communities of countries emerge. For example, in Figure 4.6.c, it is possible to identify a group consisting of Germany, Mexico, Brazil, Argentina, Netherlands, Spain and Italy: countries whose teams participated in the 2014 FIFA World Cup. Also, it is possible to observe edges among Malaysia, Indonesia, China and Australia, reflecting the disappearance of the Malaysia Airlines flight MH370 on 2014. Those two long-term events, for instance, sparked several events in our dataset, and the interactions between the protagonist countries are reflected in our analysis.

We further explored trends of co-protagonism by analyzing the similarity of countries over time. Given two countries, we computed their Jaccard similarity based on the events of a time window of one week. Figure 4.7 shows the time series between United States and Great

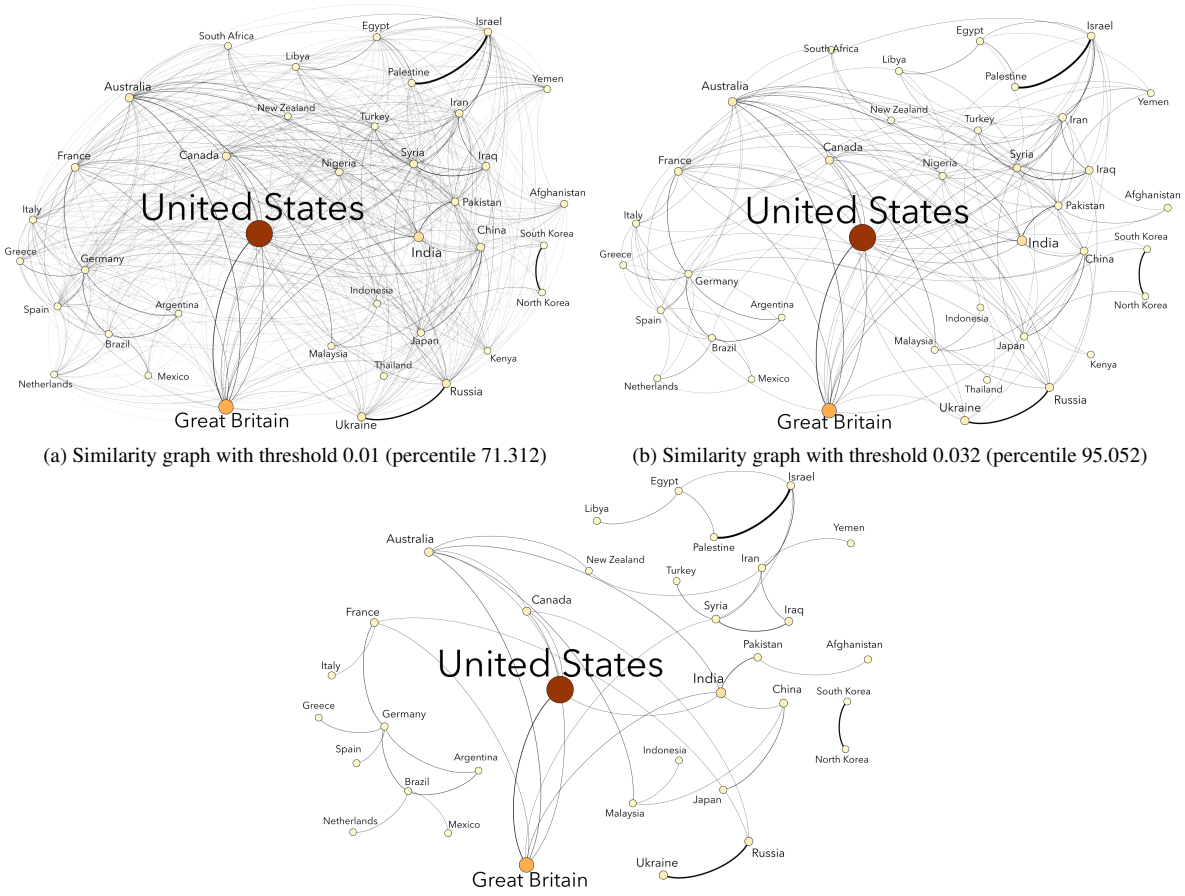
Country i	Country j	x'_{i}	x'_j	Distance
Turkey	Indonesia	198	172	1.1442
Yemen	Turkey	202	198	1.3416
Afghanistan	Turkey	323	198	1.5304
Libya	Turkey	253	198	1.6050
Egypt	Palestine	316	360	1.6496
Malaysia	Turkey	262	198	1.8096
Japan	Spain	354	258	1.8327
Italy	Japan	315	354	1.9018
Brazil	Spain	236	258	1.9060
Germany	Pakistan	371	453	2.0674
Israel	Syria	561	647	2.4463
Russia	Ukraine	823	921	2.5557
Nigeria	Pakistan	412	453	2.5822
Canada	China	715	646	2.6025
Iran	Syria	496	647	2.6838
Iraq	Iran	654	496	2.9270
France	Canada	627	715	3.7859
Australia	France	974	627	4.1398
India	Australia	1,561	974	4.8339
Great Britain	India	4,015	1,561	41.7719

Table 4.2: Pairs of countries that had the closest **pi** vectors according to the Euclidean Distance. x'_{i} is the number of events in which country i was a protagonist.

Britain, Malaysia and Australia, and Russia and Ukraine. Each of those pairs of countries showed different characteristics in terms of how their similarity evolved over time. The US and Great Britain did not show notorious bursts of similarity over time, although they had high overall Jaccard similarity (Table 4.1), showing that although they were co-protagonists in several events, there was not a particular situation that suddenly increased their similarity in a narrow time span. On the other hand, Malaysia and Australia showed a burst starting in March 2014, shortly after the disappearance of the Malaysia Airlines flight MH370 (similar patterns arose when inspecting the relationship with Indonesia and China). Finally, Russia and Ukraine showed high values of similarity over time, starting roughly in December 2013 and those patterns were maintained throughout 2014. This scenario correlates well with the case study reported in Section 5.3.1.

Another aspect that we explored was the interest that different countries had in events that occured in different geographical regions. In other words, we explored the *protagonist-interest* relationship between countries. To do this, we represented each country c_{i} as its corresponding $\mathbf{pi}(c_{\mathrm{i}})$ vector (Equation 4.2).

We adjusted the original representation of the protagonist-interest vectors (Equation 4.2)



(a) Similarity graph with threshold 0.01 (percentile 71.312)

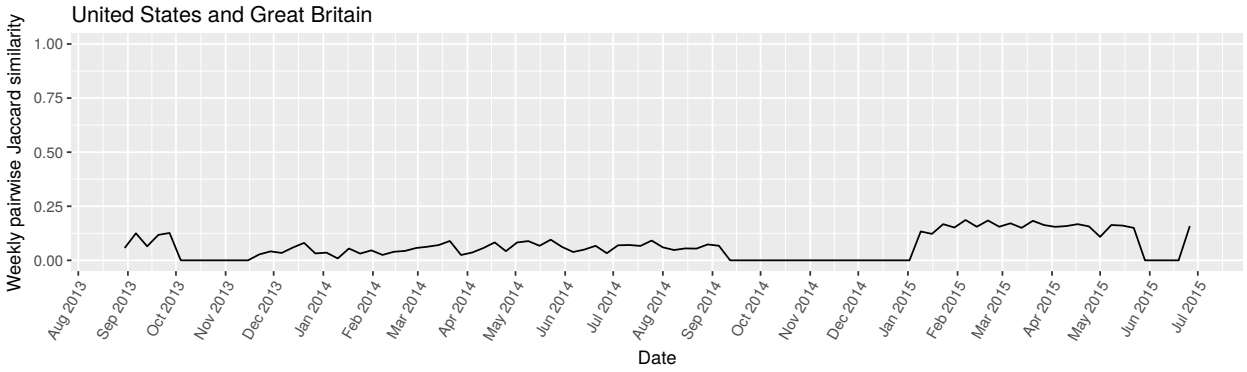
(b) Similarity graph with threshold 0.032 (percentile 95.052)

(c) Similarity graph with threshold 0.06 (percentile 98.309)

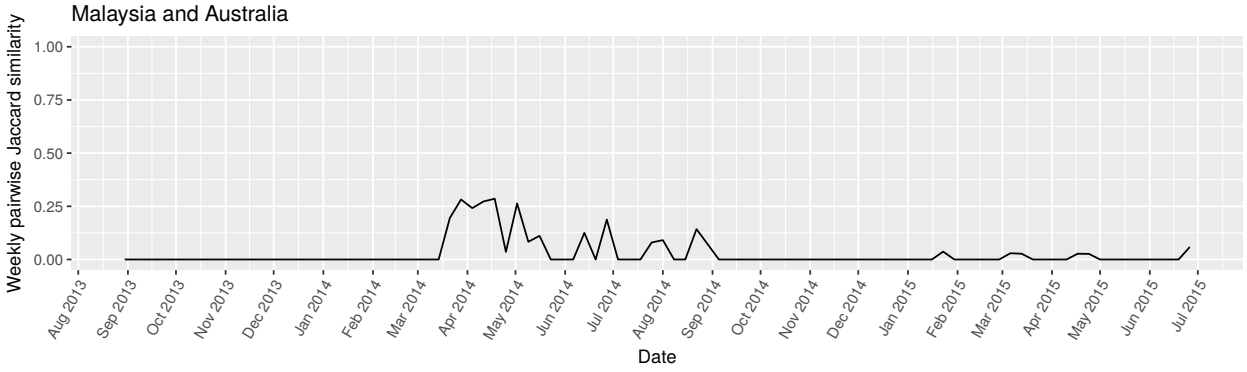
Figure 4.6: Similarity graphs of countries using the Jaccard similarity as the weight for the edges. Each node is a country and an edge between two nodes corresponds to the Jaccard similarity between those two countries. An edge is present if the similarity is higher than the given threshold. The node size and color represents the number of events in which each country was a protagonist, and the thickness of and edge represents the similarity.

in order to mitigate the data bias, which was reflected in that some countries were overly represented, because they produced much more tweets than others (Figure 4.5.a). Hence, instead of counting the number of events with c_j as a protagonist, for which c_{i} expressed interest, we prefered to measure the interest of c_{i} in c_j as the difference between the average number of events of other countries in which c_{i} was interested, with respect to the number of events of c_j in which c_{i} was interested. In other words, our original interest measure was normalized by the average interest shown by c_{i} in other countries. Using this new interest measure we applied Euclidean distance to find the country c_2 with the closest **pi** vector to another country c_1 (Table 4.2). Given that there were countries that expressed interest in only a few events, or that were protagonists themselves of very few events, we only report the countries that were protagonists of at least 167 events (i.e., the average number protagonist events per country in our dataset).

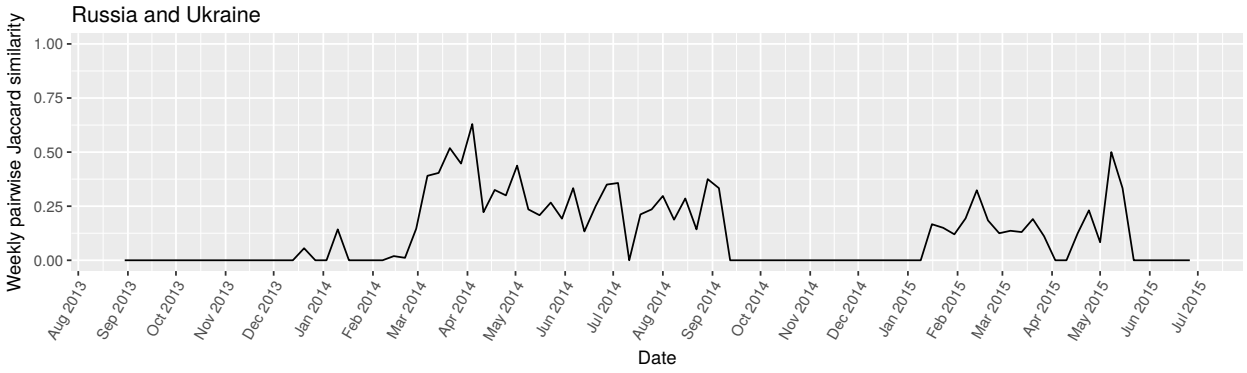
We observed that Turkey had strong ties with other countries, being very close with several other countries according to protagonist-interest relations, such as Indonesia, Yemen,



(a) United States and Great Britain.



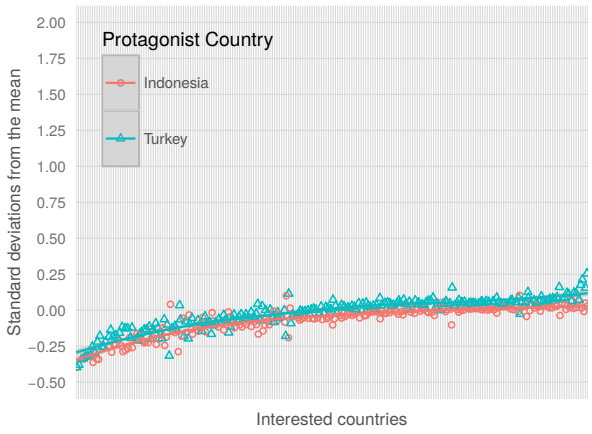
(b) Malaysia and Australia



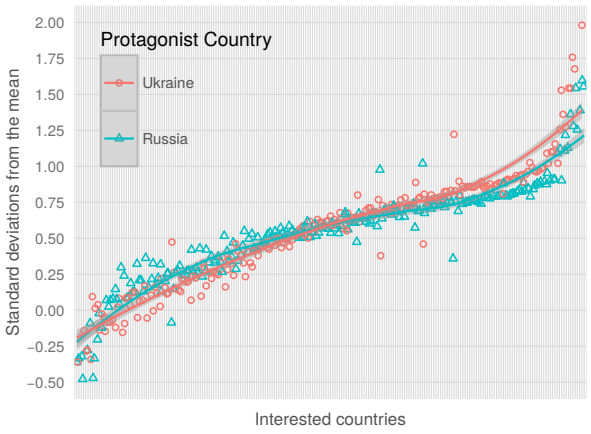
(c) Russia and Ukraine.

Figure 4.7: Timeseries of the Jaccard similarity between co-protagonist vectors of selected pairs of countries over time. The value of similarity is computed for all the events in the given week. Data from October 2014 and December 2014 was not available.

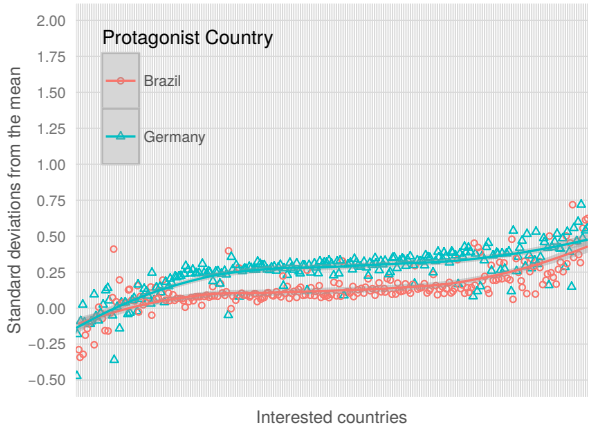
Afghanistan, Libya, and Malaysia. Furthermore, other similar countries were Italy and Japan, Brazil and Spain (and also Brazil and Germany); these similarities are be explained by the events triggered in the 2014 FIFA World Cup. Notably, Russia and Ukraine standout again, showing not only that they were protagonists of roughly the same events, but also that they were seen with similar interest by the rest of the world, making the impact that the Ukrainian crisis had on the news more evident. We also noted that most of these countries are close geographically, and as well as other countries, mostly from Asia. We argue that these results are another sign of the bias in our dataset: the perspective of international news as seen by English-speaking countries.



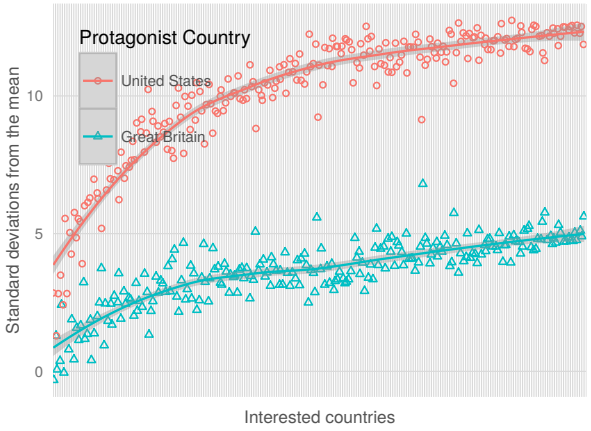
(a) Indonesia and Turkey.



(b) Ukraine and Russia.



(c) Brazil and Germany.



(d) USA and Great Britain.

Figure 4.8: Protagonist-interest plots for selected countries. Each plot shows the level of interest (y-axis) displayed by the other counties of the world (listed along the x-axis) in the events of the featured pair of "protagonist countries". Country labels in the x-axis have been omitted for readability purposes.

Table 4.3: Events with most international impact, measured as the number of countries which showed interest higher than the 99-th percentile of overall interest.

Event Description	Tweets	Users	Outlets	Countries
Death of actor Robin Williams (2014)	1.8M	1.3M	48	202
FIFA World Cup final between Germany and Argentina (2014)	494K	385K	40	144
FIFA World Cup starts (2014)	476K	358K	45	143
Super Bowl starts (2015)	1.1M	849K	35	130
New Year's Eve (2013)	325K	279K	31	127
Soccer Player Luis Suarez is suspended from World Cup (2014)	213K	157K	38	106
Charlie Hebdo shooting in Paris (2015)	629K	328K	50	102
Grammy Awards (2015)	682K	432K	31	97
Boxing match between Mayweather and Pacquiao (2015)	779K	522K	37	97

Finally, we explored events with the highest impact, considering international and local

Table 4.4: Events with most local impact, measured as the number of tweets coming from events with only one interested country, whose interest is higher than the 99-th percentile of overall interest. All events happened on 2015.

Event Description	Tweets	Distinct Users	Outlets
US Supreme Court ruled in favor of same-sex marriage	51K	50K	7
Delhi Legislative Assembly election	35K	13K	3
Labour party said it will scrap the non-domiciled tax status	32K	15K	10
Tornado strikes Texas	31K	6K	4
TV appearance of Delhi chief minister candidate Arvind Kejriwal	30K	10K	1
Hillary Clinton announces presidential bid	30K	30K	3
Football player Cardale Jones announces he is returning to school	28K	22K	3

events. For this analysis, we considered all international events (regional and global). We counted the number of different interested locations for each event, however only considering interest measurements within the 99-th percentile of the dataset. From this analysis we were able to observe that the events with the highest overall impact covered several topics, and that the most recurrent events were sports and entertainment. Events like the death of the actor Robin Williams caused the most international impact, with a large number of tweets from 202 countries. This was followed by sports events, such as the 2014 FIFA World Cup, the 2013 Super Bowl and the boxing match between Floyd Mayweather and Manny Pacquiao (Table 4.3). Other events with high impact included New Year's Eve for 2013, the *Charlie Hebdo* shooting in Paris, and the Grammy Awards in 2015. We also observed that the coverage of different news outlets was higher for these events. On the other hand, events with local impact consisted mostly of political events, such as political elections and debates, with the exception of a natural disaster and a sports event. We observed that in this case the coverage of different news sources was lower in relation to high impact international events, as well as the number of tweets involved.

4.3 Known Limitations

The main limitation of this exploratory analysis is the event extraction methodology we used to collect events from Twitter. Indeed, the news event extraction methodology relies on the headlines published by news media accounts, which provides good precision in terms of reporting events that did in fact exist in the real-world, but might omit informative events that did not receive media coverage. Therefore, the current data extraction approach can fail to retrieve events such as citizen movements and other important events that were informed only via social networks. In addition, in the current data extraction setup the initial seeds for the event collection came from a reduced list of news media accounts, with limited country coverage and languages. Although the news event dataset likely represents a great majority of the news events and related tweets posted on Twitter, the collection will miss the long tail of events that had impact in other less represented countries worldwide. Another relevant aspect to consider about bias is the demographics of Twitter users. Some studies have shown that the population of Twitter users is biased towards certain characteristics. For example Mislove

et al. [130] analyzed a sample of users from the United States and concluded that Twitter users were mainly men and were non a random sample in terms of race or ethnic groups. Other researchers have reached similar conclusions regarding the unbalanced population of Twitter users from other countries, like Italy [179] and Great Britain [125]. Nevertheless, our event representation allowed us to observe political relationships among countries that happened in the physical world by only analyzing social media data.

Finally, we note that although our proposed event representation can be considered generalizable to other social media platforms, we have not validated it on other sources of information besides Twitter. It is not certain that for other social media platforms we will have enough information, regarding user location and data availability, in order to produce accurate event representations.

4.4 Summary

In this chapter we presented a spatio-temporal context-aware representation for news events in social media In particular we introduced two types of geographical contexts for events: 1) protagonist locations, and 2) interested locations. The first corresponds to locations, in this case geopolitical divisions, that were involved in the event itself, and the second corresponds to locations where the event's information had the most impact. By considering both contexts, we defined two types of geographical scope for news events: provenance and impact.

In addition, we presented an exploratory analysis with our event representation used on almost two years of news events collected from Twitter. With this analysis we observed that our event representation allows us to study the geographical features of a news event from two complementary perspectives that merge the physical world and its reflection in social media. We also observed that this representation allowed us to study how international relationships develop and how they evolve over time.

Chapter 5

Visualization of News Events by their Geo-temporal Representation

As expressed earlier, understanding patterns can be impossible without the collaboration of human and data mining techniques. Indeed, computer-based data visualization provides means for people to view and interact with data that can lead to abstract interpretations that could not be possible with computers alone. As stated by Munzner [131]:

> *visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.*

Visualization of social media data has been extensively studied in recent years, due to the huge volume of data produced in these platforms by users. The interest produced by the events portrayed in social media has given rise to a large body of visualization research on how to portray events from different perspectives. For example, several works in the literature focus on the representation of events as their most representative keywords and how they evolve over time using river metaphors [200, 116] or other kinds of visual summaries [16]. Other approaches use more than one perspective at a time, including multiples views like TwitInfo [123] that incorporates overall text sentiment, shared photos like the work of Dörk et al. [54] and videos that elicit conversation in Twitter like in the work of Diakopoulos et al. [49]. When considering the geographical aspects of an event, several works focus on either showing the geographical distribution of messages over a map [123, 84, 106] or the locations where the event occurred, like Event Registry [58]. SensePlace2 [121] is one of the few examples that includes both the geographical distribution of messages and locations of the event. Nevertheless, as far as we are aware of, none of the exiting tools include the political links generated by the interaction of locations in the physical world.

In this chapter we present Galean, our prototype of a visual interface to explore and retrieve news events based on our proposed spatio-temporal context-aware event representation. Given that our model is focus on understanding the relationships between events and locations, it does allow users to explore international relationships among countries. We present our system's interface and high-level architecture. We show the usefulness of our tool by presenting two case studies, and by evaluating its effectiveness for new Information

Retrieval tasks, such as: retrieving events that have particular countries as protagonists, and following international relations among countries over time.

5.1 Interface Design

Galean's interface design is based on the Visual Information-Seeking Mantra: Overview first, zoom and filter, then details-on-demand [162]. Its interface (Figure 5.1) is composed of three main components: (i) filters and search (Figure 5.1.a, top section); (ii) a list of events and the main map (b and c in the middle section of Figure 5.1); and (iii) the timeline (Figure 5.1.d, at the bottom). A video demonstration of this tool is available at `https://vimeo.com/150260355`. In addition, a prototype of Galean focused only on Chilean news is available at `http://galean.cl`. In the future, the international version of Galean will be made available in the same location as the Chilean version.

Next, we describe the interface and its components in detail.

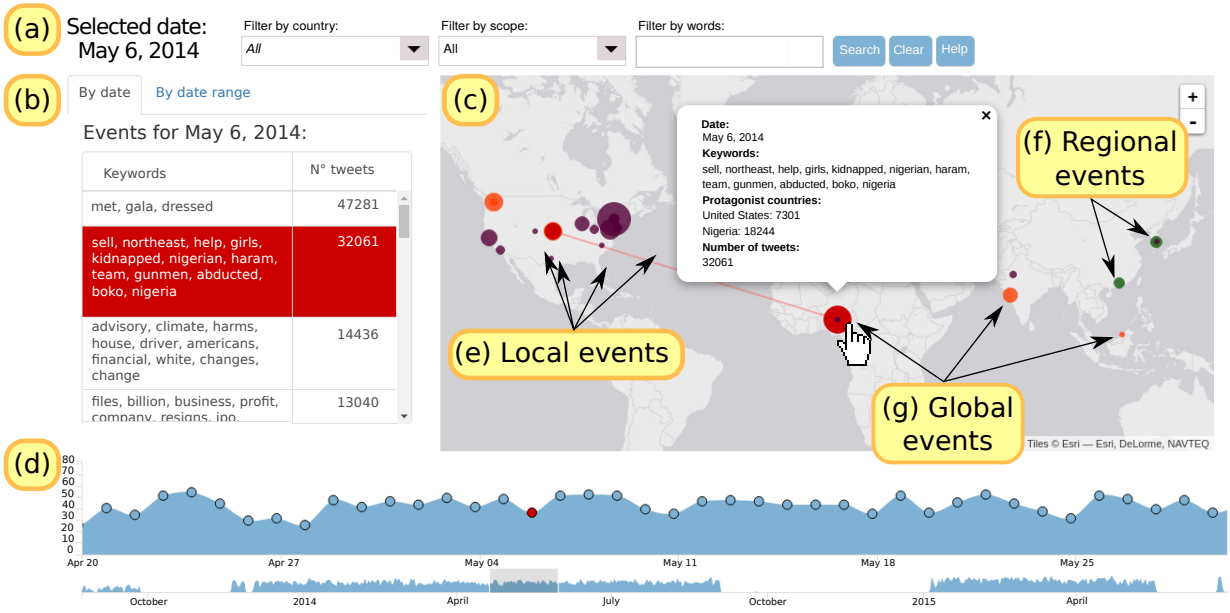


Figure 5.1: Galean overview. (a) Filters and keyword search options are in the top section. In the middle section, (b) a list of events by date and date range, and (c) the main map. (d) The timeline at the bottom shows the volume of news events over time. (e), (f) and (g) indicate local, regional and global events respectively.

Overview first: Main map and timeline. The main map, table of news events and timeline provide a simple overview of thousands of tweets about news events. The main map shows events in their geopolitical context, represented as bubbles placed over the country or countries of their provenance. If the event is located in a particular city within the country, the bubble is placed in the city. On the other hand, if only country level information is available for the event, the bubble is placed on the country's capital. The size of each bubble represents the relevance of the event, measured by the volume of tweets associated

with it. Purple bubbles (Figure 5.1.e) represent events that are of local provenance (i.e., events in which only one country is the protagonist). Green bubbles (Figure 5.1.f) represent regional provenance (i.e., more than one country is involved in the event, but all of them correspond to the same continent). Orange bubbles (Figure 5.1.g) represent events which are of global provenance (i.e., more than one country is involved in the event and they belong to at least two different continents). If the cursor is placed over a bubble, a pop-up appears with information about the event. When the cursor is placed over green or orange bubbles – regional or global events– links appear to indicate the other countries that are relevant for that event. For example, in Figure 5.1 we observe several local events in the United States in May 6, 2014, indicated by purple bubbles located on this geographical area. In particular, the event with the highest impact is located in the West Coast. Some regional events (green bubbles) are located in South Korea and Brunei, and some global events (orange bubbles) are located in India, China and the United States. In addition, we highlight a global event that links the United States and Nigeria, which corresponds to the United States' intentions to send aid to Nigeria in response to the kidnapping of a large group of schoolgirls claimed by Boko Haram, in 2014 [194].

To the left of the main map, the interface contains a list of events displayed by their most representative keywords and number of tweets. The timeline at the bottom shows the overall distribution of events over time, providing a historical overview of events per date. It is built as a focus-plus-context component of all the news events from the database. If a date is selected, the main map is updated showing only the events of that day. The map and timeline were implemented using Leaflet [109] and D3.js [45], respectively.

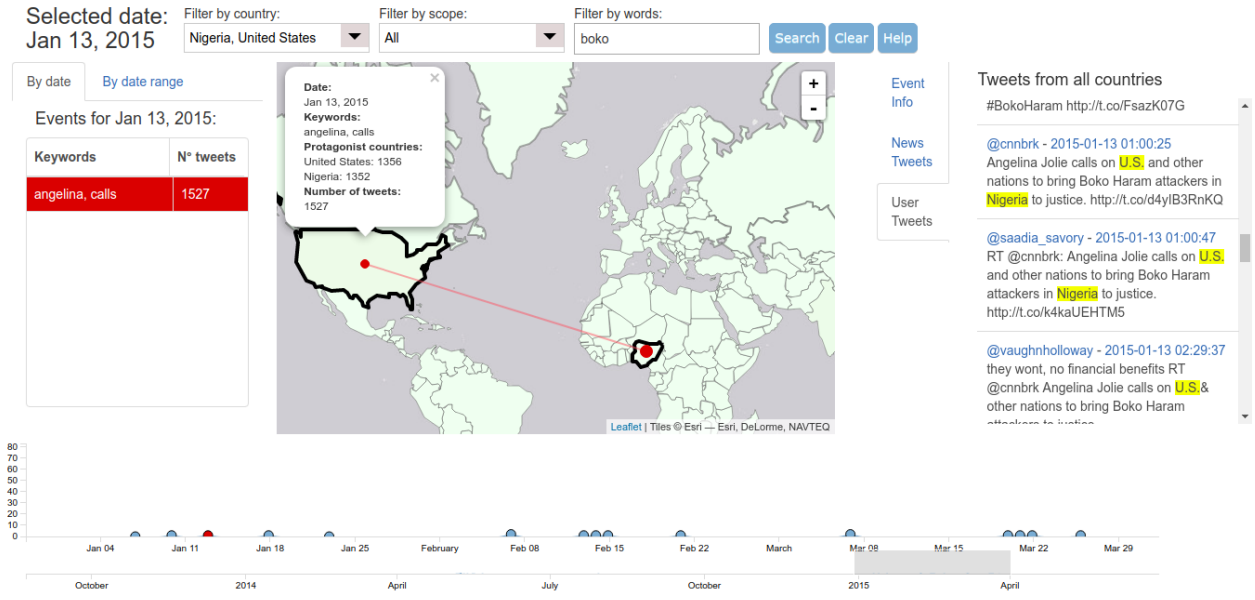


Figure 5.2: Galean interface after applying filters on protagonists countries and keywords. It retrieves and displays events related to the kidnapping of Nigerian schoolgirls by the Boko Haram terrorist group.

Zoom and filter. If the top filters of the interface are applied, the map, the list of events and the timeline are updated according to these filters. Events can be filtered by (i) whether they have one or more protagonist country, (ii) the scope of their provenance (local, regional

or global, defined in Chapter 4), and/or (iii) by keywords. In particular, if more than one protagonist country is selected then the system retrieves only events in which those countries interact. For example, we can explore how the relationship between the United States and Nigeria evolved over time, based on the schoolgirls kidnapping by selecting both countries in the country filter and the word `boko` in the search box. After applying the filters, the timeline only shows events that meet the requirements imposed by them. By manually inspecting some dates in the timeline, we can retrieve several events related to that topic. For instance, in Figure 5.2, Galean shows in its timeline which events satisfy these criteria, particularly displaying a related event on January 13, 2015.

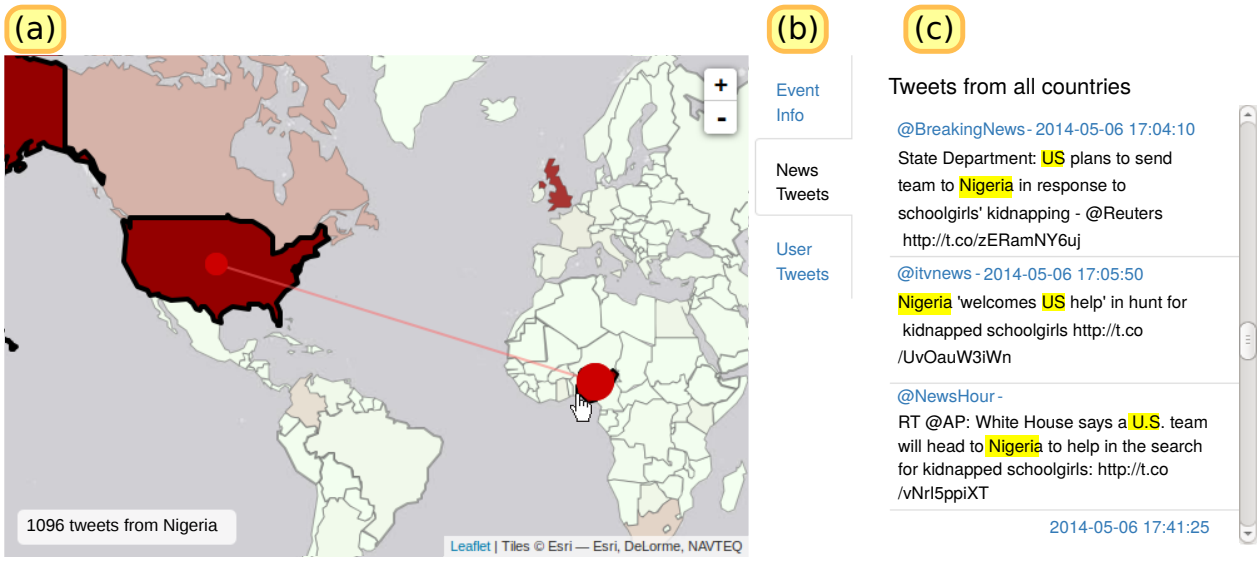


Figure 5.3: Details on demand for the news event about the intentions of the U.S. to send aid to Nigeria during the schoolgirls kidnapping (May 6, 2014). It shows the (a) geographical distribution of tweets, (b) additional information of the news topic divided into categories, and (c) tweets related to the event.

Details on demand: selecting a news event. To inspect a particular news event in depth, the user can click on its corresponding bubble in the map or on the list of events that is displayed. When an event is selected, shown in Figure 5.3, the map is updated to show a choropleth of the geographical distribution of tweets according to the countries that display interest in the event (countries from which users post tweets about the event). The event's protagonist countries are highlighted with a darker outline. Additional information for the event can be found at the right-hand side of the map. This information consists of a general event summary and of event tweets, categorized by source (i.e., regular Twitter accounts or news outlet accounts), shown in Figure 5.3.b. By selecting these different sources, users can view a set of headlines for the event (i.e., when selecting news outlet tweets), or compare the people's perspective against that of the media. Finally, if a country is selected from the choropleth, tweets will be filtered to show only those coming from the selected country in chronological order.

In particular, Figure 5.3 shows that most of the tweets related to the schoolgirls kidnapping come from the United States, Nigeria, Canada and Great Britain. In particular, the tweets shown in Figure 5.3 reflect the media's reaction to the event.

It is important to mention that our event exploration tool does not provide event ranking nor tweet ranking functionalities at the moment. The tool displays all of the events that match the user-defined spatio-temporal filters, and tweets are listed in chronological order. Ranking at the moment is not within the scope of our work, but it could be interesting to address in future versions.

5.2 System Architecture

We present a general overview of the architecture for generating our event representations in order to use them in our application. The architecture, shown in Figure 5.4, consists of the following three parts: *"input"*, *"event representation generator"*, and *"visualization"*. The first component, (1) "input", is not part of the contributions of this dissertation, and is currently fulfilled by using existing methods, which can be replaced transparently as long as the requirements detailed next are met. On the other hand, the other two components, (2) "event representation generator" and (3) "visualization", are the core of our contribution and therefore essential to our system.

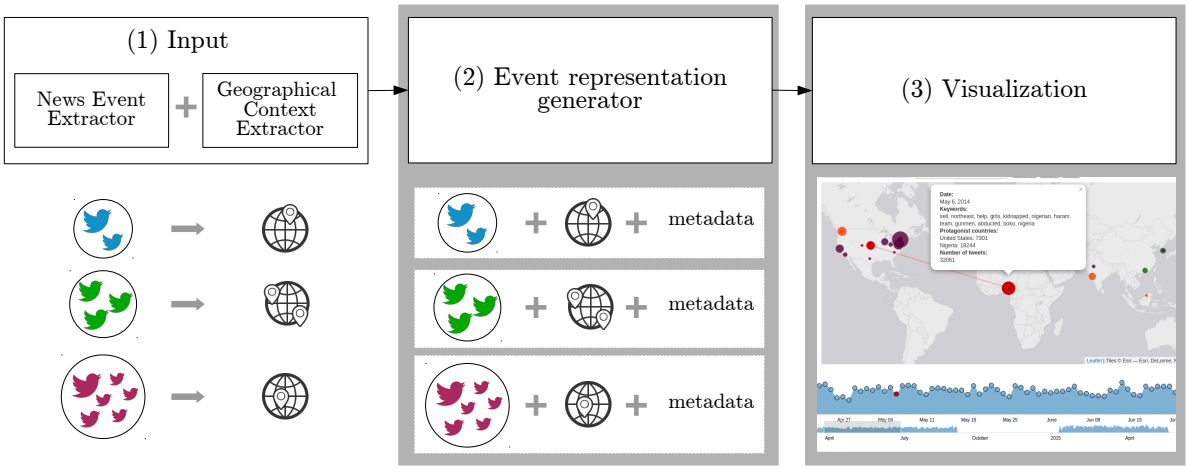


Figure 5.4: Framework consisting of three parts: 1) input, which collects data related to news event activity from social media and extracts its geographical information; 2) the event representation generator, which generates our representation of the input events and 3) the visualization, which consumes these events. Our contribution is related to the two latter modules, the first module can be replaced according to the task and/or state-of-the-art.

Given an input from the Twitter data stream we specify the following components of our framework (the particular setup for our proposed applications is detailed in Section 4.2.1):

1. **Input:** This module requires two subparts, the *"news event extractor"* and the *"geographical context extractor"*.

   (a) *News event extractor:* This submodule must output groups of tweets, where each group of tweets T should represent a cohesive news topic E. In particular, most

of the tweets in the set T of an event E must be on the topic of a particular news events. However, as we use a high-level representation of events, some noise is tolerated (i.e., tweets that do not correspond to the event).

(b) *Geographical context extractor:* This submodule associates spatial context to each tweet in T of each event E produced by the "news event extractor" module. Therefore, it must provide the geographical locations of the places mentioned in the text of the message and the geographical location of the author of the message (i.e., protagonist and interested locations, respectively). This module must locate the majority of the tweets in E correctly (i.e., with good precision) based on GPS coordinates and/or textual content, so that locations mentioned in tweets can be geotagged, and users can be geotagged as well (users can set their location using GPS coordinates or by using natural text).

2. **Event representation generator**: This component creates the event representations E for each of the groups of tweets provided by the "input" module. In particular this module must create a tuple E for each event, as specified by our definition in Section 4.1. This means that it has to produce the date D of the first tweet, a set of keywords K that describe the event, the set T of tweets and the $\mathbf{P}$ and $\mathbf{I}$ location vectors of the event.

3. **Visualization**: This module consumes the event representations produced by the "event representation generator" module and produces the event visualization interface.

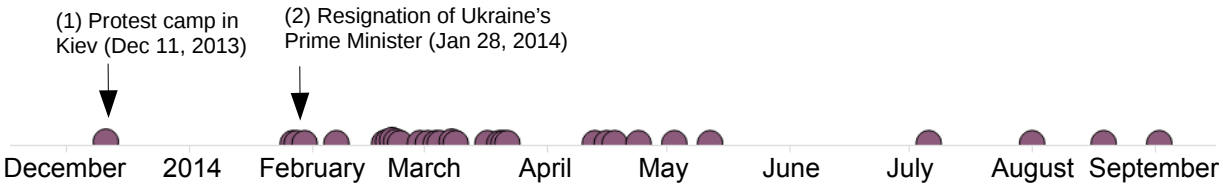
The backend of the tool was implemented with Flask [154], the python based microframework, and frontend with Javascript. The code was modularized so each component behavior was as independent of others as much as possible.

5.3 Tool Validation

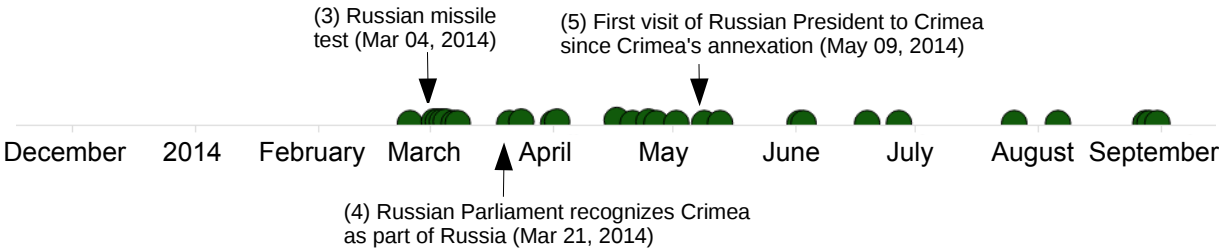
In this section we present the validation of our tool with case studies, a qualitative study with experts and a user study. For the evaluation of Galean we used the same dataset described in Chapter 4, section 4.2.1. In particular, the processes of *News Events Extraction* and *Geographical Context Extractor* described there are the same subcomponents of the *input* component described in Section 5.2. As previously mentioned, although the input data is important for the outcome of the final application, we consider the event detection and extraction to be beyond the scope of this current thesis. In practice, this means that the way in which events are extracted can be replaced by another methodology in the future.

5.3.1 Case Studies

We use Galean to explore two selected news events: the *Ukrainian crisis*, dating approximately from November 2013 until today, and the *earthquake in Nepal* in April, 2015.



(a) Timeline of **local events** from mid December 2013, to late September of 2014. Local events started to emerge in December, 2013 and then decreased in number and frequency around March, 2014.



(b) Timeline of **regional events** from mid December 2013, to late September of 2014. Regional events started in March, 2014.

(c) Timeline of **international events** from mid December 2013, to late September of 2014. International events started to emerge around March, 2014.

Figure 5.5: Timeline of (a) local, (b) regional, and (c) international events in the Ukrainian crisis between December, 2013 and September, 2014. Russia and the United States were the external countries that became the most involved in the Ukrainian crisis according to our analysis.

Ukrainian crisis. This event corresponds to the long-term conflict in Ukraine, which consensually started in November 2013 when the Ukrainian government decreed to suspend signing the "Association Agreement" [11] with the European Union. We used Galean to discover events related to the Ukrainian crisis, by selecting `Ukraine` in the country filter and the term `crisis` in the keyword filter. This retrieved only events that occurred in Ukraine and that contained social media messages with the term `crisis` between November, 2013 and March, 2015. To understand how local, regional and global events differ, we used Galean's filters to select the scope of each event. At the beginning (December, 2013), the majority of events were of local scope (Figure 5.5.a), meaning that Ukraine was the only protagonist country at that time. Months later (March, 2014), regional and global events started to appear, indicating that other countries became involved in the crisis (Figure 5.5.b and Figure 5.5.c), tendency that started to decrease later in May, 2014. More precisely, Galean displayed 36 regional events about the Ukrainian crisis, 28 with Ukraine and Russia as protagonist countries. On the other hand, we found 48 global events, 12 of them involving only Ukraine and the United States, and 31 of them involving Ukraine, Russia and the United States.

In addition, we tracked some local events, in particular those related to the evolution of the protests in Kiev [164], and their consequences, such as the resignation of the Ukrainian Prime Minister at that time [201] (both events are marked in Figure 5.5.a as `(1)` and `(2)` respectively).

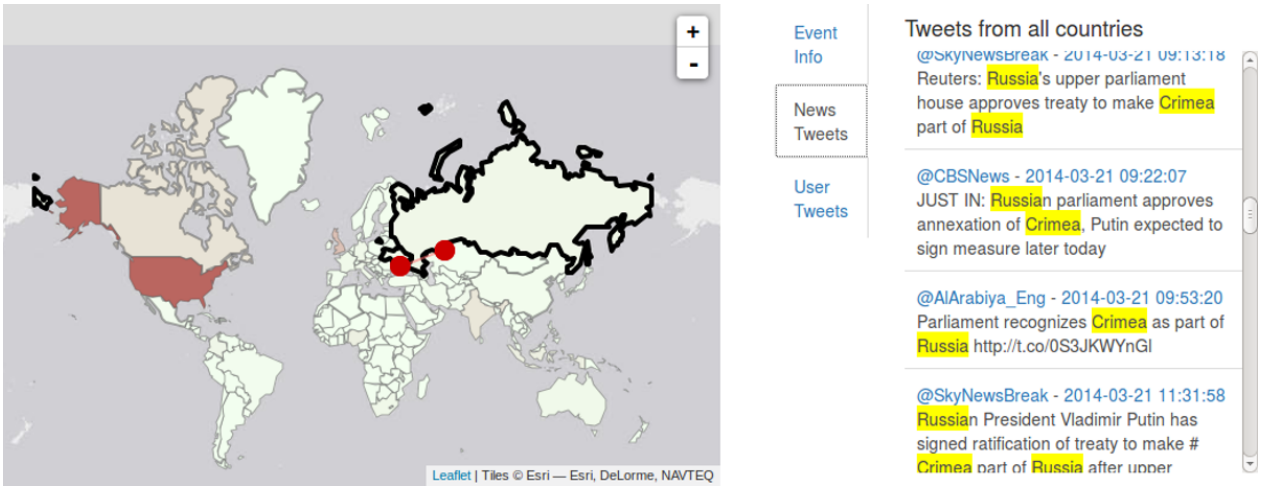


Figure 5.6: Russian Parliament recognizes Crimea as part of Russia (Point `(4)` in Figure 5.5.b). Event detected on March 21, 2014. Total number of tweets: 7,660.

According to Galean, Russia and the United States were important actors in the Ukrainian crisis. Therefore, we explored more in detail some regional and global events. We found a series of events that related Russia to Ukraine, for example, when the Russian Parliament recognized Crimea as part of Russia in March 21, 2014 (Figure 5.6), and when, as consequence, the President of Russia, Vladimir Putin, celebrated Victory Day during his first visit to Crimea in May 9, 2014. Both events showed a strong impact on Twitter, with 7,660∼ and 11,647∼ related tweets.

Events that involved the United States included sanctions towards Russia [60], or accusations about Russia sending military help to separatists in Ukraine [173]. We used the filters provided by Galean to find relevant protagonist countries for certain events, and to track these events in time. For this case study, we observed an overall tendency of international, regional, and global scope events producing a greater impact, than local scope events.

2015 Earthquake in Nepal. In this second case study we show how Galean can help users in crisis management, by looking at the causes of certain events. The starting point of this search was a news event about Japan signing an agreement to provide a loan for Nepal's earthquake recovery programs in December 2015 [172]. We retrieved events related to the earthquake by selecting `Nepal` as a protagonist country and `earthquake` as a keyword filter. In terms of scope, we obtained 24 local events, 7 regional events and 18 global events.

Regarding the earthquake's impact in social media, we observed that people's interest grew as the event evolved as evidenced by an increase in the number of related tweets and also of countries from which users displayed interest. In addition, we found emerging relationships between other countries, besides Nepal, such as the United States and India, as a consequence of having provided external aid for aftershocks.

Given that our dataset extends only up to June, 2015, we were not able to follow the complete lifecycle of this event. The last global event related to the earthquake in Nepal stored in our database was from May 16, 2015, which accounted for the recovery of the bodies of the crew of the U.S. Marine chopper that went missing while helping victims [171].

Figure 5.7: Geographical distribution and sample tweets about the donor's conclave for the reconstruction in Nepal. Event detected on June 25, 2015. Total number of tweets: 2,565.

However, after clearing the keyword filter in order to use only the filter by country, we found a regional event in June 25, 2015 about a donor's event among several countries to rebuild Nepal [181]. This event had Nepal and India as protagonists because the biggest donation came from India (Figure 5.7). Another agreement of this particular event was a loan from Japan to Nepal, which actually corresponds to the initial news that started this case study. Hence, by starting from that news, which is consequence of a past crisis situation, we were able to track its origin and subsequent events.

5.3.2 Expert Feedback on the Visual Tool

We conducted a qualitative study of Galean with six domain experts using Pair Analytics [17]. Two specific aspects were investigated: (i) how intuitive and easy the tool was to use, and (ii) whether the tool could be used for the experts' day-to-day work in long-term news analysis. It is important to note that for this study our prototype only implemented two categories for provenance: local and international (regional was added afterwards). The international category included regional and global events. For the scope of the study, we consider long term analysis following the development of a news event for more than one month. Its design and results are presented in this section.

Design. Six users (two men and four women) participated in the study. They were recruited by e-mail or by online social network messages. They are journalists (4 people), Information and Management Control Engineer (1 person), and Computer Engineer (1 person). The ones who are not journalists work on news analysis as part of their daily work. Their ages ranged from 25 to 35 years old. They were not economically compensated and participated voluntarily in the experiment. The experiment was conducted in an office in the Department of Computer Science of the University of Chile.

Initially, the experimenter described the objectives and the procedure of the study to the participants. Then she described the tool and its components, allowing the participants to

use the tool as s/he needed to explore its features. The participants are required to carry out three small tasks to get used to the tool's interface and exploration capabilities. Later, they were required to perform four more tasks which were designed for the analysis of news events which required the users to be more focused. Finally, they were asked to discuss their experience using the tool. All the interactions of the participants with the tool, the conversation with the experimenter and their comments were recorded. The participant did not have to write any answer or comment by themselves. To extract the answers of the participants, the experimenter transcribed the recorded session.

Two major aspects were investigated: (i) to discuss how intuitive and easy the use of the tool was, and (ii) to consider whether the tool could be useful for their daily work.

Experimental Scenario and tasks. The data set described in Section 4.2.1 was used as the scenario of the experiment. The events asked to be analyzed in this section were manually selected by following two criteria: (i) they seemed to be easy to understand, and (ii) they were probably not known by the participant.

As mentioned earlier, two types of tasks were selected to evaluate the usability of the tool. The first three tasks were selected in order to get used to the tool. In particular, to study if the timeline was intuitive enough (`T1`), if a particular filter was understood by users (`T2`), and if it was understood that filters could be combined to make up more complex queries (`T3`). The first three tasks are described as follows:

- `T1:` Name the date when most events happened. Select it and name a couple of news events.
- `T2:` Filter the events in that date by scope: local and international. Name which scope comprises most news events.
- `T3:` Filter the events in that date as being local and from the United States. Name the event with most tweets which discuss it.

The following describes the four investigative tasks. The first two (`T4` and `T5`) focus on the exploratory capabilities of Galean and how it presents the development of a news event over time. The last two (`T6` and `T7`) focus on discovering pattern behavior of news events and their propagation on Twitter. The patterns requested were directly related to the underlying model. For example, we wanted to explore if participants observed that with a particular scope have similar behavior in terms of impact or propagation.

- `T4:` Search for the news related to the Crimea crisis in 2014 and briefly describe the events found. In particular, focus their development over time, the number of tweets and the geographical distribution of tweets discussing the events.
- `T5:` Search for news events that are related to the Washington mudslide in 2014 and conduct a similar report.
- `T6:` Explore local and international news events and compare them. Describe how each category behaves over time and how much impact on social media they had, measured

in volume and geographical distribution of tweets.

- T7: Select international events regarding United States and Iraq and describe how they evolve over time. Repeat the process with United States and Chile. Compare both scenarios and explain their differences.

Results. The results of the expert feedback are reported from three perspectives: usability, usefulness, and patterns found in the data.

Usability: Tool design and completion of tasks Overall, all of the participants were able to complete all the tasks without substantial problems. They understood the main interface and how to use it. The participants seemed happy and entertained to work with the tool. They also seemed to be interested in what it showed to them. In fact, it was common to hear comments such as "this is fun". Most of them agreed that with more practice the tool would be easy to use.

The issues regarding the interface design correspond to the confusion about the date filters, and the clutter provoked by several events being shown up in the same location. There were minor difficulties at interpreting the maps and the difference between a tweet and an event in the main interface. Nevertheless, all participants were able to understand the source of the data and how it was displayed.

At the beginning, it was not clear for some participants that when selecting a date from the timeline, the date filter also updates automatically. In addition, participants frequently asked for the possibility to select a date range.

The second issue to address was the overlapping of event bubbles on the main map. Even if strategies were applied to overcome this problem, sometimes it was hard for participants to distinguish between news events when several bubbles were on top of the same country on the map. Participants also suggested that the pop-ups would have information about the country in which a bubble is, as their geography knowledge was not always perfect.

Regarding the data set used, participants declared that the news events displayed by the tool corresponded to their general knowledge of the news. However, Twitter is a noisy source of data, and as a consequence, it was possible to find news events that were mixed with other news. Also, Galean showed that some countries participated in news that did not exist in the real world.

Usefulness: Galean for daily work All participants agreed that the tool is useful for analyzing news. However, most of the journalists indicated that its usefulness for their daily work will depend of the type of analysis they need to conduct. For them, it was very important to know the source of information and Galean did not provide it at the time. Because of that, the question "who tweeted this?" was commonly asked during the study. They explained that it is very important for them to know if the source of data was a news media or a common Twitter user. This is because messages from common users are not always useful, as it is hard to distinguish whether they were a rumor or a fact.

Participants also easily distinguished news that developed over time and occasional events. They were able to follow events over time and track them as a consequence of a starting news event. For example, when performing the fifth task (T5), some participants not only found events about the mudslide in Washington, but also the end of the search of victims. Regarding Galean, one participant stated: "[The tool] is very useful for conflict analysis. It will be great to use this tool for the case of the boy killed in Ferguson [59]. [...] Because the first time might be an accident, but the third is thuggery."

On the other hand, participants seemed interested in exploring links between countries given international events. In the particular analysis of the Crimea crisis (T4), some participants were surprised to find some events were linked with other countries, such as the United States. When this happened, they were interested in exploring when and how the other countries got involved in the conflict. While conducting these tasks, one participant pointed at the timeline and stated: "...for someone who doesn't know there was a war in Ukraine, it will be interesting to see that at some point everybody started talking about it."

Patterns in the data Participants found that international events had more impact on Twitter than local events. In particular, the news events with the greatest impact were those in which countries such as the United States, Russia and Great Britain were involved. P5 stated: "Something remarkable and evident is that international events are inclined to involve particular countries and local events are not". Participants also commented that even if the distribution of local events was more homogeneous than international events, local events from influential countries would have greater impact on Twitter. A participant said: "Well, the United States will always be under the magnifying glass [...] Remember that boy killed in Ferguson? [59] Everyone knew about that. But internationally, who remembers the boy killed by a policeman in Peñalolén [Santiago, Chile]?" It is important to note that these patterns might have been consequence of the dataset seeds used in the empirical setup. Nevertheless, participants were able to recognize these patterns using Galean and they were not surprised to find them.

One participant did not trust the tool completely regarding this last point. Even if she was able to see that there was a relationship between the interest of some countries and the impact of the news on Twitter, it was not enough to derive conclusions from the event for her. She commented: "Without knowing the importance the news had in media, I can't say how much impact it had in Twitter."

5.3.3 User Study

We conducted a more general user study to obtain evidence of users' perception of the visual tool, in relation to its efficiency and effectiveness for retrieving information about international relationships based on news reported on Twitter. As in the expert feedback evaluation section, we only divided events into local and international provenance scopes. Our user study was designed to test these two main hypotheses:

Hypothesis I: Users will retrieve information about relationships between countries within the context of news events in a more efficient and effective way using Galean than a competitive baseline interface.

- **H1.1** Users will take less time to complete analysis of the news event.
- **H1.2** Users will retrieve more relevant information about the countries involved in a real world news event.

Hypothesis II: Users will have a better subjective perception of Galean than a competitive baseline interface.

- **H2.1** Users will have a lower cognitive load to complete a news event analysis.
- **H2.2** User will perceive Galean as a better system to conduct news event analysis.

Study design. The study had a within-subjects design, in which participants had to analyze news events using Galean, as well as using a competitive baseline interface. The baseline is shown in Figure 5.8 and described next. At the beginning of the session, the goals of the study were described to the participants and they were asked to fill a pre-study survey with demographic information (Appendix A.2). Next, the study was divided into two stages of news event analysis, each of them requiring participants to use one of the interfaces. At the start of each stage, participants followed a brief tutorial of the assigned interface and were given indications on how to complete the task. Participants started to complete the first event analysis only once they declared to understand the interface and the task. After they had finished, they were asked to fill the NASA Task Load Index [82] and a post-study survey (Appendix A.3). Once they were ready, subjects repeated the same procedure, with a different news event, with the second interface. We selected two news events for the users to analyze, and then asked questions about them such as "*when did the event happened?*" or "*which countries were involved in the event?*". The complete list of questions are in Appendix A.1. The events to analyze were: (i) the news about Saudi Arabia leads airstrikes on Yemen rebels, on March 26, 2015 [41]; and (ii) the news about Malaysia airlines lost contact with flight MH370 on March 8th, 2014 [169]. The answer to each question were submitted in a web form included in the web interface where they have to conduct the news analysis. To prevent a learning effect, we counterbalanced the order of presentation of each interface and of each event. In addition, the interface only gave access to tweets of one news event at a time.

With the two interfaces to test in addition to the two events to analyze, we divided the participants in four groups (Table 5.1).

All evaluations were conducted using the Chromium Web Browser in computers with an Intel Core i5 CPU, 8GB of RAM and Ubuntu 14.04 installed. Participants spent close to one hour to complete the whole study.

Baseline. We built the baseline based on SensePlace2 [121], shown in (Figure 5.8). We chose

Group	Interface 1	Event	Interface 2	Event
A	Galean	Airstrikes	Baseline	MH370
B	Galean	MH370	Baseline	Airstrikes
C	Baseline	Airstrikes	Galean	MH370
D	Baseline	MH370	Galean	Airstrikes

Table 5.1: Study design conditions.

this tool as the most similar to ours in terms of the geographical information displayed. In the upper part, users were able to search by date and keywords. In the bottom-left, users could read tweets that matched the search. On the right side, the interface displayed geographical information in a similar fashion than SensePlace2, in which a map showed the number of tweets published by country and the geographical entities found in the content of the tweets. Since our focus is at country level, the geographical distribution of the tweets was not displayed as a grid, but only as a choropleth. The geographical entities found in tweets are represented as bubbles located in the geographical coordinates of the location. Both, the country area and the bubble on the map, could be used as filters.

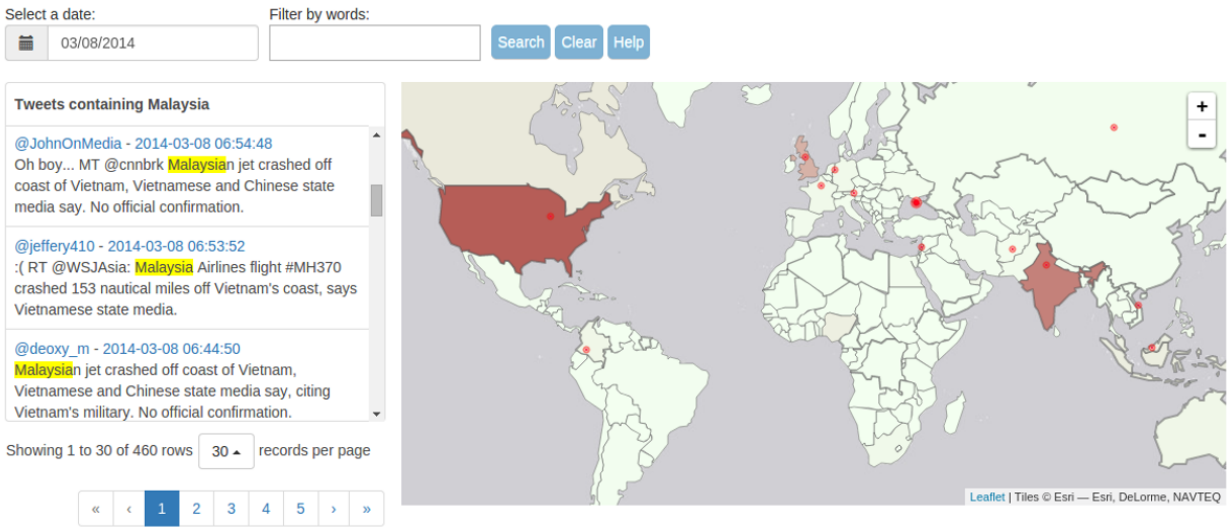


Figure 5.8: Baseline interface. The top section shows search by date and by text options. At the bottom left, it displays the tweets that matched the user search. Finally, at the bottom right it displays a map with the geographical distribution of tweets as a choropleth and the geographical entities that appear in the content of the tweets as bubbles over the resolved location. Both the choropleth and the bubbles representing a location can be used as filters for the tweets.

Participants. Participants were recruited by e-mail and online forums in the Engineering School of the University of Chile. Given that our dataset was in English, we required them to have a good level of non-technical English. From the total of 30 participants recruited (3 of them were women), 5 of them were less than 20-years old, 20 were between 21 and 30-years, and 5 were between 31 and 40-years. In addition, 10 of them were undergraduate students, 8 were Masters students, and 12 were PhD students. Participants were not economically

Quantitative metric for efficiency	**Galean**	**Baseline**	***p*-value**
1. Overall time to complete the task	895.58 ± 57.10 secs.	955.65 ± 74.82 secs.	0.18
2. Overall time when Baseline was second	954.62 ± 82.90 secs.	784.62 ± 55.84 secs.	$*$**0.013** (d = 0.6)
3. Overall time when Galean was second	832.60 ± 77.68 secs.	1138.07 ± 128.93 secs.	0.00051 (d = 0.74)
Quantitative metric for effectiveness			
5. Precision to retrieve countries involved in the event	0.95 ± 0.02	0.87 ± 0.04	0.062
6. Recall to retrieve countries involved in the event	0.36 ± 0.04	0.35 ± 0.04	0.4

Table 5.2: Objective metrics to evaluate Galean efficiency and effectiveness to retrieve international relationships among countries within the context of a news event. The p-value was obtained with paired 1-tailed t-test.

compensated, however refreshments were available during the study.

Results. Our study only partially supported hypothesis **H1**, evaluated by objective behavioral metrics of efficiency and effectiveness, but it completely supported **H2**, assessed by users' perception on the tasks performed during the study.

H1. *Objective measures of efficiency and effectiveness:* In terms of efficiency, users spent less time to complete the task using Galean ($M = 895.58$, $SD = 317.9$) than using the baseline interface ($M = 955.65$, $SD = 416.57$), though this difference is not significant ($p = 0.18$). We argue that a reason for this difference being not significant is a learning effect, since some key components on the interfaces to complete the task were similar between conditions, such as the search box, the map, and the list of tweets. Therefore, we investigated this possible learning effect, and observed that users indeed spent less time using the second interface, but that this difference was more pronounced when Galean was second. By comparing the difference in time when Galean was the second interface ($p < 0.001$, Cohen's d = 0.74) versus when the baseline interface was used second ($p = 0.013$, Cohen's d = 0.6), we observed that the effect was larger when Galean was second. This result is interesting because Galean had additional components and interactions to learn from, which indicates that Galean was more efficient for this task than our baseline.

Regarding effectiveness, there was no clear difference in *recall* between Galean ($M = 0.36$, $SD = 0.2$) and the baseline ($M = 0.35$, $SD = 0.2$), $p = 0.4$, when used for retrieving countries. In terms of *precision*, Galean obtained a better performance ($M = 0.952$, $SD = 0.11$) than the baseline ($M = 0.871$, $SD = 0.24$), $p = 0.062$ when retrieving countries involved in a news event, although this difference was barely non-significant. A summary of all objective metrics are in Table 5.2.

NASA Task Load Index (21 gradations scale)		
Question	**Galean**	**Baseline**
1. Mental demand	8.29 ± 0.82	9.84 ± 0.71
2. Physical demand	4.13 ± 0.72	5.19 ± 0.79
3. Temporal demand	8.16 ± 0.70	8.58 ± 0.64
4. Performance	8.19 ± 0.88	8.55 ± 0.78
5. Effort	9.39 ± 0.84	11.32 ± 0.78
6. Frustration	$^*\mathbf{5.74 \pm 0.79}$	9.03 ± 0.98
Post survey (Likert scale)		
1. How intuitive did you find the interface?	3.74 ± 0.15	3.39 ± 0.22
2. Would you use it to analyze news events?	$^*\mathbf{3.55 \pm 0.21}$	2.71 ± 0.22
3. How confident were you in the information displayed?	$^*\mathbf{4.06 \pm 0.13}$	3.39 ± 0.19
4. Did you lose notion of time while conducting the task?	3.48 ± 0.21	3.16 ± 0.22
5. Would you recommend the tool?	$^*\mathbf{3.97 \pm 0.17}$	2.87 ± 0.21
6. How much satisfied are you with the tool?	$^*\mathbf{3.71 \pm 0.17}$	2.52 ± 0.19
7. How much information do you think the interface did not allowed you to see?	$^*\mathbf{2.48 \pm 0.17}$	3.06 ± 0.17

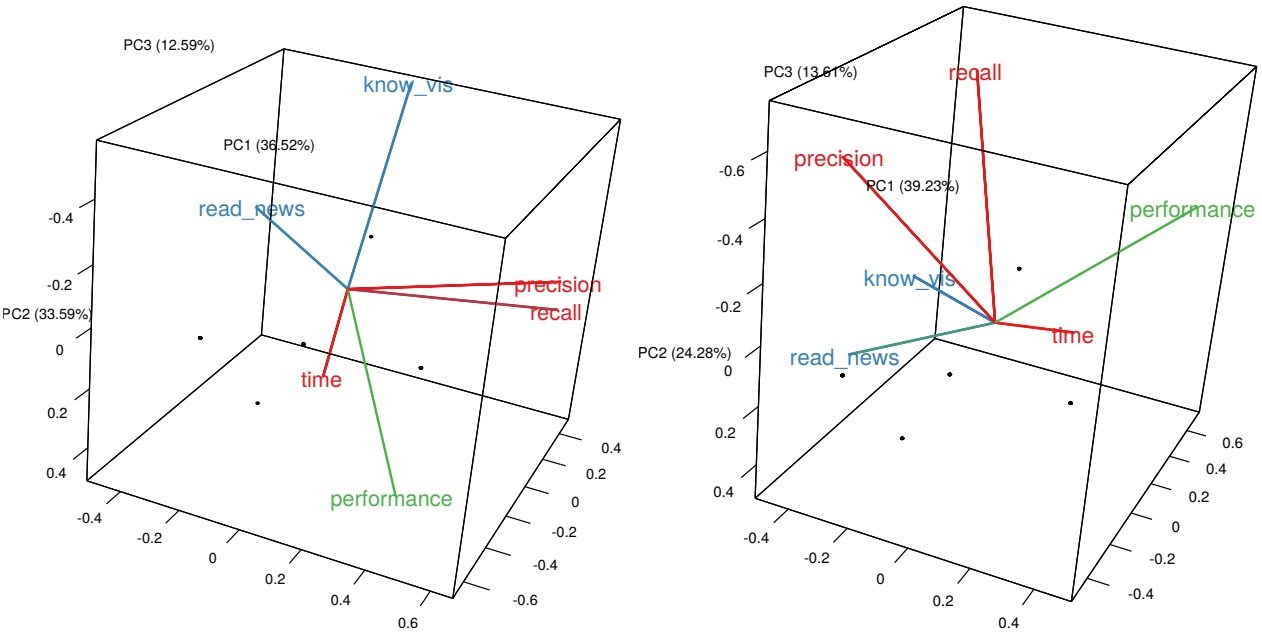
Table 5.3: Subjective metrics to evaluate users perception of Galean to analyze news events. **(* indicates p-value < 0.05, obtained with paired 1-tailed t-test)**

H2. *Subjects' perception on the interfaces.* Our study supported hypothesis H2, indicating that Galean was perceived in general as better than the baseline by users. We obtained subjective metrics by applying the NASA Task Load Index [82] and a post-study survey. Participants also showed the trend of requiring less effort to complete the task and less frustration ($p < 0.05$) when using Galean. However, participants did not register too much difference in engagement during the execution of the task. In addition, we were not able to measure engagement in the long term as it was not the goal of the current study. On the other hand, Galean was not perceived to be more intuitive ($M = 3.74$, $SD = 0.86$) to be used than the baseline ($M = 3.39$, $SD = 1.20$). As stated before, this could be because Galean interface contains several components, interactions and visual metaphors that could not be intuitive at first. With respect to the final post-study survey administered with a Likert 1-5 scale, people felt more confident about the information displayed in Galean than in the baseline ($p < 0.05$), they showed greater satisfaction ($p < 0.05$) and they were more likely to recommend it for eventual analysis of news events ($p < 0.05$). A summary of subjective metrics are in Table 5.3.

User Agreement. In order to measure the level of agreement between users' perception over Galean versus the baseline interface, we used the Intraclass Correlation Coefficient (ICC) [163]. We calculated ICC between users (raters) over post-study survey questions (samples)

and we report and interpret the values using the guidelines described by Koo et al. [100]. Values of ICC less than 0.5 are indicative of poor agreement, between 0.5 and 0.9 indicate moderate to good agreement, greater than 0.9 indicate excellent agreement.

The ICC results show a moderate to good level of agreement between users. For the case of Galean, the level of agreement was good (ICC = 0.887) with a 95% confidence interval from 0.722 to 0.977 ($F(6, 210) = 8.88$, $p < 0.001$). For the case of the baseline interface, the average measured ICC was moderate (ICC = 0.723) with a 95% confidence interval from 0.317 to 0.943 ($F(6, 210) = 3.61$, $p = 0.002$). ICC estimates and their 95% confident intervals were calculated using the *irr* package[1] version 0.84 within the R statistical package version 3.3.1 based on a mean-rating ($k = 31$), absolute-agreement and 2-way random-effects model.



(a) PCA Biplot for Galean

(b) PCA biplot for baseline interface.

Figure 5.9: 3D biplots of the principal component analysis for objective and subjective metrics for both interfaces. Some metrics were removed for clarity. Subjects knowledge such as how familiar were participants with visualizations (know_vis) and how frequently they read news (read_news) are in blue. Objective metrics of time, precision and recall are in red. Subjective metrics of user's perceived performance are in green. While precision and recall are related to previous user knowledge on the baseline, loading on the same direction of PC1, in Galean precision and recall are more related to user's perceived performance and barely related to previous user knowledge.

Discussion. Our results show that in terms of user perception metrics, Galean clearly outperformed the baseline, but in terms of objective performance metrics, Galean shows only a tendency of better efficiency and effectiveness than the baseline.

To investigate these results further we conducted a principal components analysis (PCA) to integrate both the objective and subjective metrics (Figure 5.9) and we analyzed them by means of a biplot. A biplot is a projection-based graphical display which allows us to

[1] `https://cran.r-project.org/package=irr`

analyze multivariate data [65]. The word "bi" refers to the joint display of both rows and columns of an original data matrix, which has been projected into a lower rank approximation with rank $n = 2$ (2D biplot) or $n = 3$ (3D biplot). In our case, rows are user subjects and columns are variables, such as precision, recall, or time spent on an interface. We obtain the rank two and rank three approximations of our original matrix via PCA. Biplots are used for multivariate data analysis in areas such as sociology [66], genetics [202] and bibliometrics [176]. The interpretation of biplot displays is demonstrated by Gabriel [65] and more recently by Greenacre [75]. For instance, the closer the angle between vectors in the biplot, the larger the correlation between the variables represented by the vectors.

From this analysis we highlight two main results which support this discussion. The first is that for Galean, the subjective and objective metrics of performance were more consistent than for the baseline. Indeed, we observe in Figure 5.9 that precision and recall are closer to each other (in terms of angle between the vectors) and to the question about performance in TLX for Galean than for the baseline. Secondly, in the biplot for the baseline we observe that the variables 'familiarity with visualizations' (know_vis) and 'how frequently they read news' (read_news) are closer to the vectors of precision and recall and load in the same direction of the first principal component (the horizontal axis, which accounts for the larger variance in the data), whichs might indicate that previous knowledge of the users influenced their performance rather than the interface itself, though further analysis and a user study with a larger sample size are necessary to support this claim.

In summary, the additional evidence collected with both objective and subjective metrics indicates that Galean improves over a competitive baseline in several aspects.

5.4 Interface Design Evolution

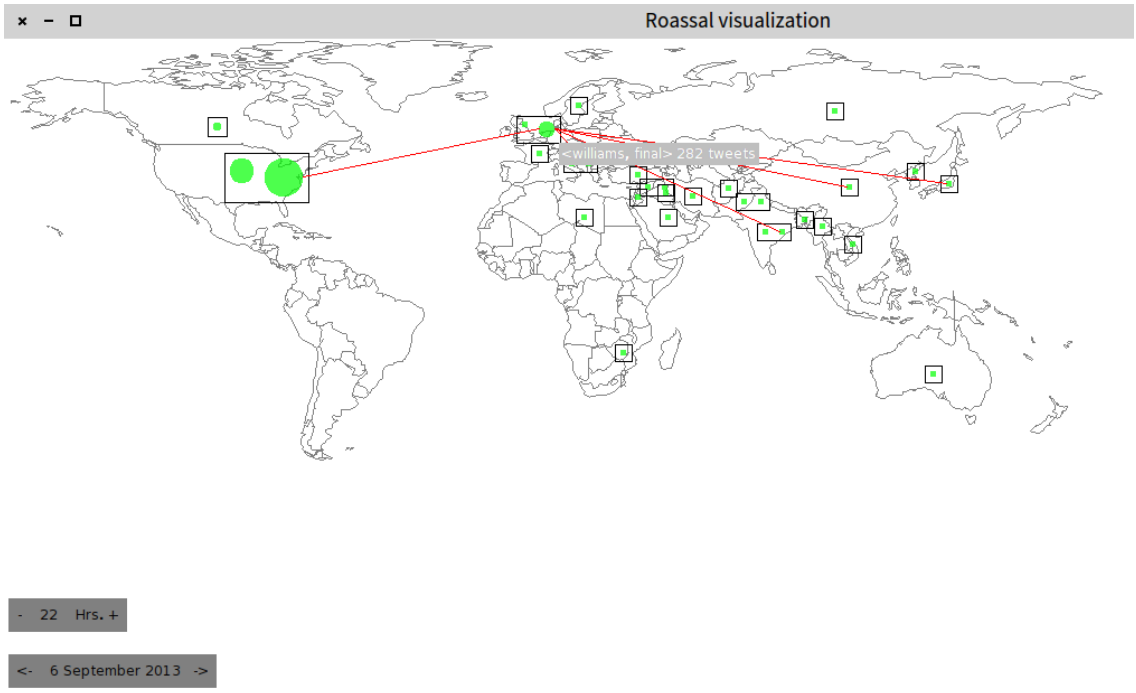
Although Galean's goal hasn't changed much from its beginning, its interface has changed significantly since it was conceived. This evolution was the result feedback from formal studies, from visualization experts or from our own requirements to obtain insight from the data. In this section we describe the evolution of its interface, detailing the reasons that motivated important changes within it. Its first design was implemented in the Pharo Smalltalk-inspired environment [146], using the Roassal visualization engine [23]. A picture of this initial prototype is in Figure 5.10. This design was the first attempt to study the protagonist and participant countries related to news events. In this interface, news events were displayed as a green circle over a map, similar to the final designs (Figure 5.10a). Events were retrieved by the hour in which they were detected and an option at the bottom of the interface allowed users to select other dates. All events in which a particular country was a protagonist were grouped in boxes situated above that country. In this first design, an event was displayed as a green circle over as many protagonists countries that it has. When moving the mouse over an event that had more than one protagonist country, red lines would appear between the selected event and its counterpart in the other countries. When using the visualization to explore the news events we realized that this was not an efficient approach, as there was repeated information and circles were cluttered when there were several of them. This changed in later versions in which a news event was represented as a unique

circle over the country with the greatest protagonism. In addition to the links between related events, when moving the mouse over a green dot, a popup appeared indicating the keywords associated with its event and the number of tweets for that event in the country for which it is situated. Finally, the visualization allowed users to view the tweets distribution for a particular news event. By right clicking in news circle the user could select "view distribution". This action opened a complementary visualization containing circles for all the countries that participated in the event, whose size represented the number of tweets related to that country. The color of the circle depended on whether if the event happened in that country, in which case the node will be green, or if it is an external country, in which case it will be red. An example of this is in Figure 5.10b.

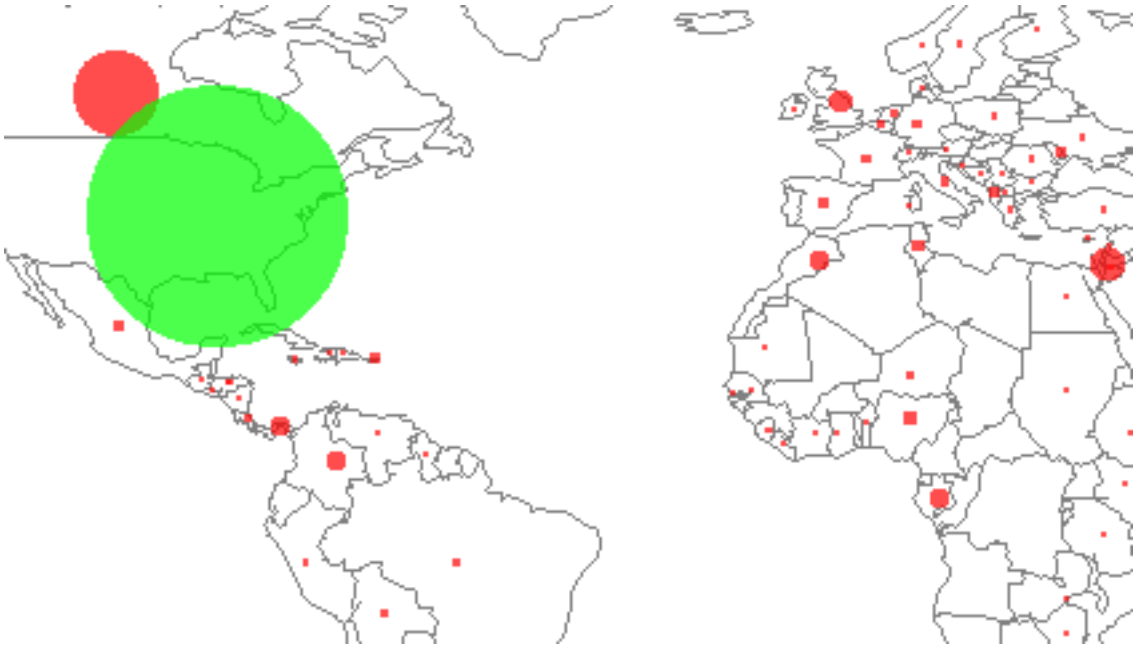
Galean's implementation was later moved to a web based implementation using web.py [167]. The main reason for this change was to implement Galean as a tool that was available to anyone who needed it without requiring one to download or install anything. An example of this reimplementation is in Figure 5.11. As observed in this picture, the main idea of a map containing the events remained as the main visualization. However, in this version we included colors to differentiate local and international events with purple and orange circles respectively. We also included other options such as search by keywords, filter by protagonist country and scope. This allowed us to more finely search and study some events in more depth. When the user clicked on an event, the interface displayed a general description of the event, a visualization of the distribution of the participant countries and a tab with the tweets text in the same interface instead of opening a news window. Similar to the previous version, the distribution of participant countries also highlighted the protagonists by using a different color. This design decision allowed us to see all the information at once and explore events more quickly. This version was presented to a group of three journalists at an informal meeting to ask them what they thought about the interface. In addition, we also asked for their thoughts about its usefulness in their daily work. We received positive feedback which motivated us to continue its implementation and the continuation of the design and the idea of working with journalists.

Before implementing the design presented in section 5.1, it passed through a intermediary version, shown in Figure 5.12. The main map remained mostly the same, except that we added a popup with detailed information of the event which was displayed where the user passed the mouse over the circle that represented it. This allowed users to have more information before selecting an event, to study it more in depth. We also maintained the filters, but distributed at the top of the interface so they would use less space. There are two important changes included in this version: a timeline that displayed the number of events per day and a table with the top 10 relevant events, measured by number of tweets commenting on them. The first component allowed us an overview of the complete dataset available. In particular, it allowed us to analyze the peaks of number of events per date for a particular long term event. The second component allowed us to have a better overview of the day being analyzed. Both features motivated us to use more powerful tools for querying the events and the tweets. We first replaced the backend framework web.py with Flask [154]. We later indexed the events in terms of date, protagonist countries and keywords using ElasticSearch [55]. This reduced the time the tool took to retrieve a particular query. Another important change was to include the identity of the author of the tweets and the division between tweets from news accounts and regular accounts. These changes were implemented

Figure 5.10: One of the first implementations and designs of Galean, developed in Pharo using the Roassal visualization engine. At the center, the interface displayed a news map with the protagonists or participant countries. To explore other events, an option was provided at the bottom of the visualization.



(a) The map displayed all the events as a green circle, enclosed by a rectangle that grouped all the events in which a country was a participant. When the mouse passed over an event with more than one protagonist country, red lines appeared to indicate this relationship.



(b) Complementary visualization that displayed the distribution of participant countries. A circle was displayed for all the countries that participated in the event, whose size represented the number of tweets related to that country. The color of the circle depended on whether the event happened in that country, in which case the node will be green, or if it is an external country, in which case it will be red.

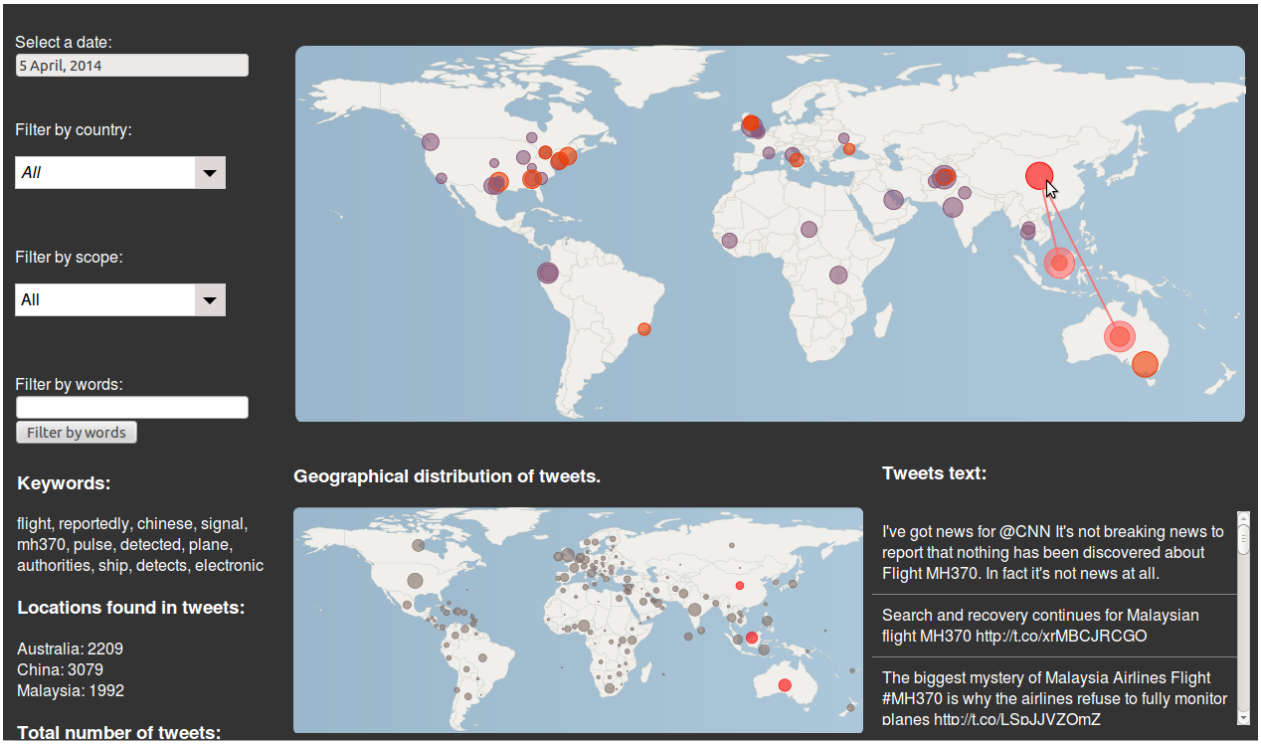


Figure 5.11: First designs of Galean interface using a web framework. We included more search and filter options and unified the visualization of protagonist and participant countries in one interface.

after the comments of participants in the qualitative study and the feedback from professor Cecilia Aragon.

Finally, the Chilean version of Galean was developed in collaboration with the National Library of Chile [2]. During this collaboration, several changes were proposed by the working team as are shown in Figure 5.13 and Figure 5.14. First, we simplify the interface so it started by only displaying the main map and a list of the events for the particular date being analyzed (Figure 5.13a). Given that this version focuses on Chilean news, we divided the map into to tabs: one displaying Chile and one displaying the world map.

In this visualization, green events are international events, purple are local events and blue are events in which one or more region of Chile are the protagonist. Chilean regions are the first-level administrative division of the country and are sixteen in total, enumerated by a Roman numeral in addition to its name. This division allowed us to analyze events that were particular to Chile and will allow us to study the centralization of news in Santiago, the Chilean capital. The lists of events displayed not only showed the most relevant keywords and number of tweets but also the protagonist countries and the most relevant headline. When there was no protagonist location found, the list indicated it with a gray icon. This allowed users a more complete overview of each event. On the other hand, all the filters and the timeline of events by date were hidden in an interactive option at the top-left hand of the interface (Figure 5.13b). The distribution of participant countries also changed as shown

[2] `http://www.bibliotecanacional.cl/sitio/`

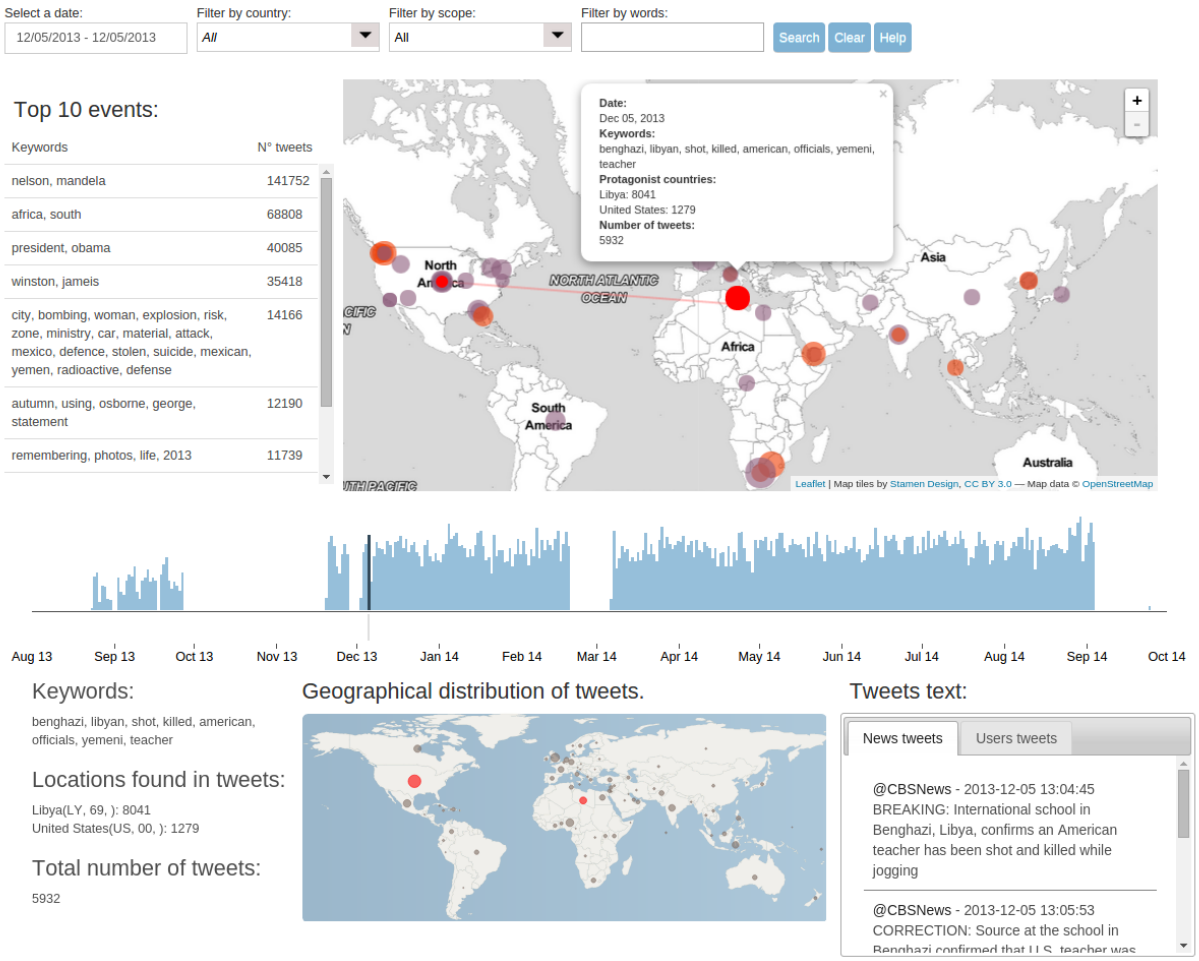


Figure 5.12: Early designs of Galean interface using a web framework and after applying the feedback received in the qualitative study. The main difference is the timeline that displayed news events per date and the list of top 10 events, measured by number of tweets commenting on them.

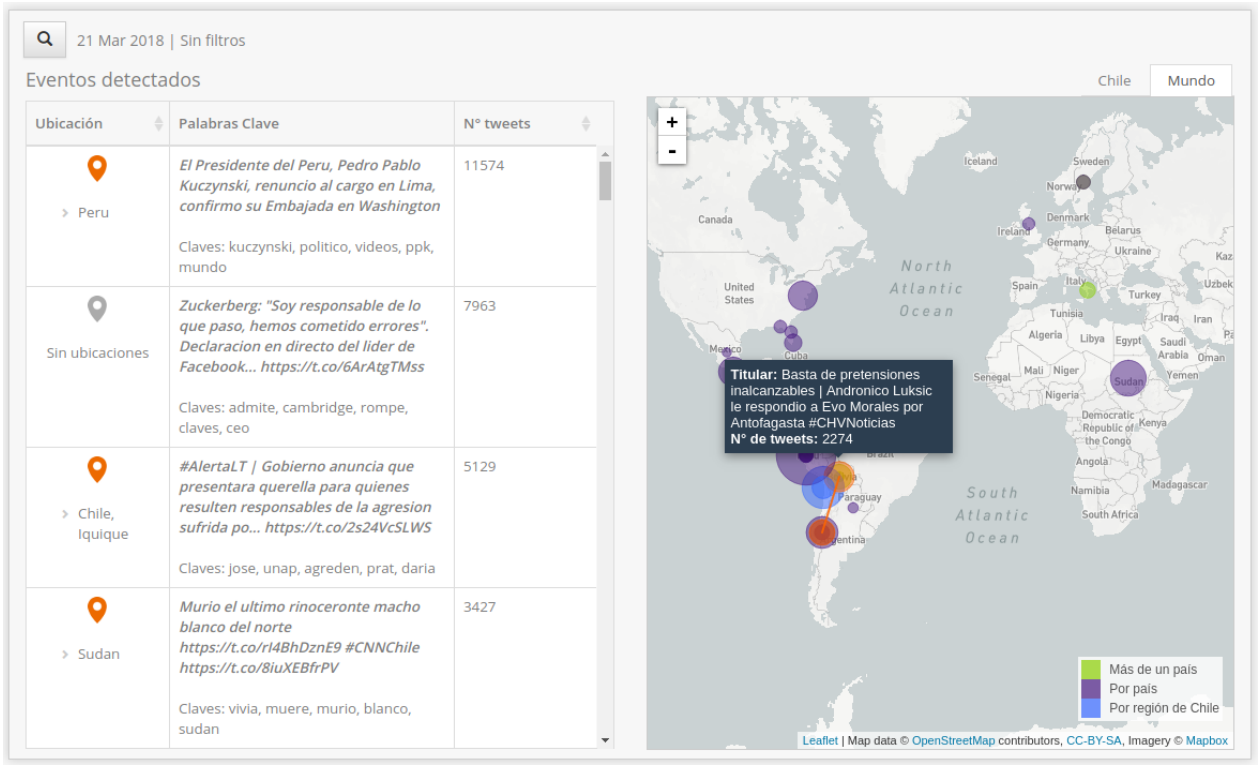
in Figure 5.14. Instead of a wider bolder to indicate which are protagonist locations in the choropeth, we added an icon over them to highlight that characteristic.

5.5 Known Limitations

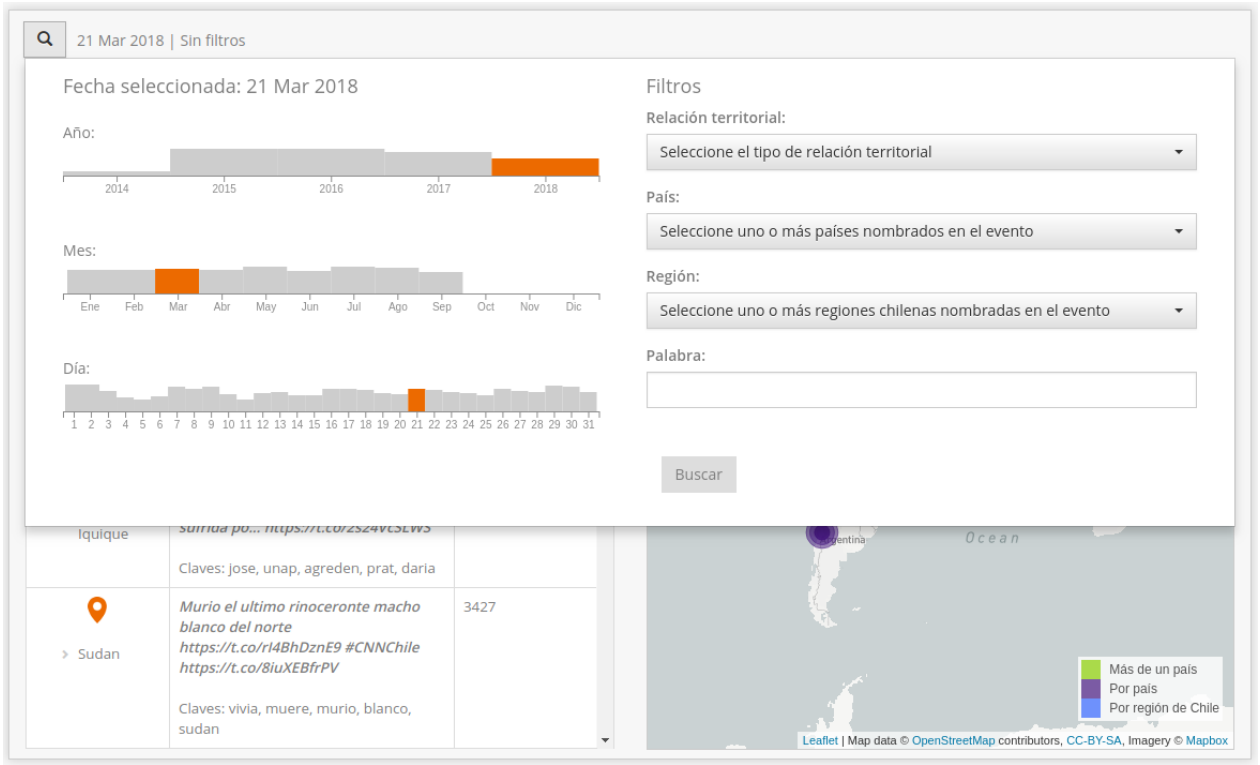
There are several limitations that we consider important to address regarding the data extraction methodology used in the empirical setup (Chapter 4, section 4.2.1), which depends on external functionalities that can impact in the validation of Galean.

Overall, basic future improvements in the input module of the architecture defined Section 5.2 should consider:

- Implementing automatic event detection techniques for Twitter based on the data stream and network properties, as well as more comprehensive microblog event ex-



(a) The filters and timeline of events are hidden in an option on the top-left hand of the interface, available when clicking on it.



(b) Filters and events timeline available.

Figure 5.13: Main map and list of news events for the available version of Galean for Chilean news events, available at `www.galean.cl`. The map contains events for local and international events, in addition to events in which one or more Chilean regions are the protagonist. The list of events contains the location of the event, if available, or a gray icon otherwise. Also, it displays the most relevant headline and the number of tweets commenting on it.

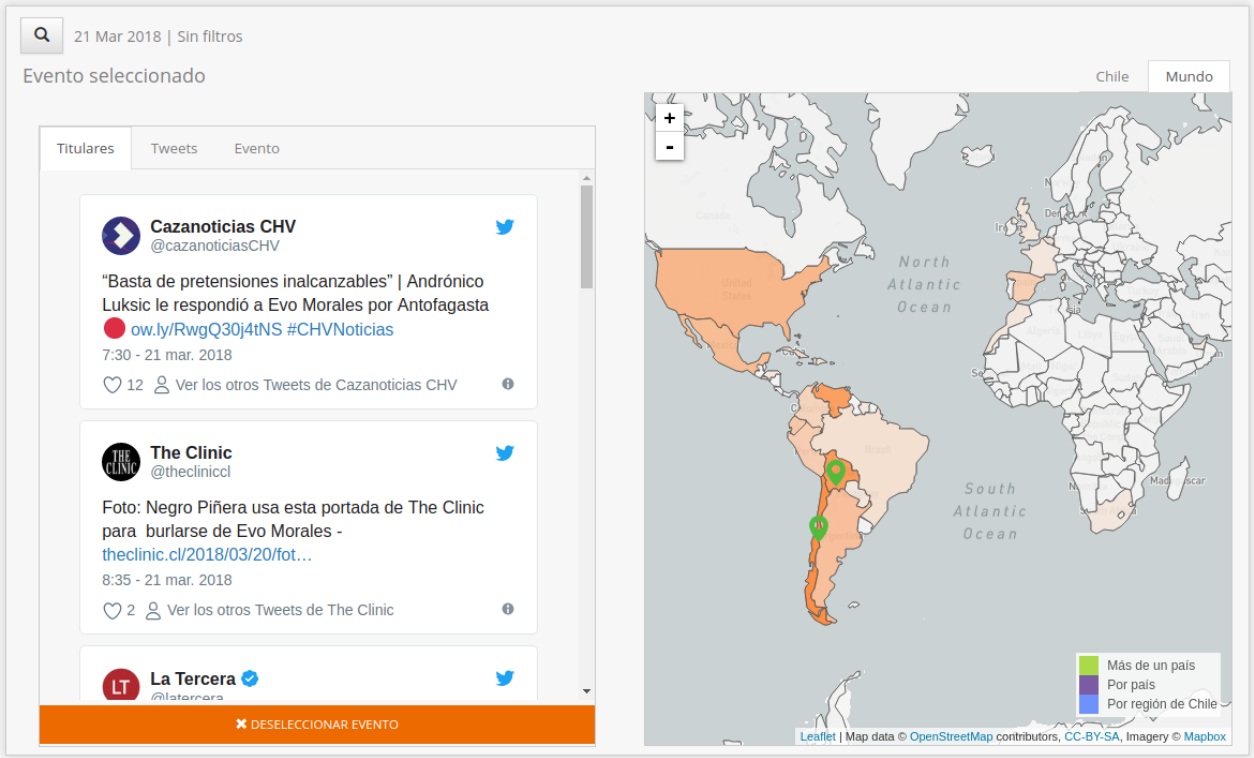


Figure 5.14: Participants distribution and tweets for a selected event.

traction approaches.

- Merging events that discuss the same news topic in different languages. Recent approaches in cross-language microblogging retrieval [73] can be integrated for news event retrieval within our framework.

- Improving the geolocation tool accuracy. Despite CLAVIN's maturity as a geolocation tool, it does not recognize location names in languages other than English (even though its documentation indicates that it does recognize alternative location names [24]).

- Adding finer granularity to the geographical context extractor of our system, in order to include more precise administrative divisions such as cities and states.

All of these improvements are however beyond the current scope of our work, which focuses on providing proof of the usefulness of the proposed event representation as well as the interactive user interface. Nevertheless, we are working on improving all of these features in future versions of our applications. For example, we have already started the task of providing more fine-grained locations for Chile and comprehensive sets of local news sources as in the work of Maldonado et al. [122].

Regarding our visualization tool, we note that even though it is an event retrieval tool, it does not focus on event ranking nor tweet ranking. At the moment the tool is centered on event exploration within spatio-temporal filters. In the future, event and tweet ranking functionalities could be added as optional features, incorporating state-of-the-art algorithms from these areas. So far, we have seen evidence that displaying the complete set of events,

and tweets, by their chronological order, appears to to be sufficient for event exploration.

5.6 Summary

In this chapter we presented Galean, our prototype of a visual interface to visualize news events in their geopolitical context. Galean is designed to allow users to manually explore news events world-wide, as well as their impact and international relations implications. Using this tool, we show that the proposed event representation allows us to perform historical analysis of events and countries over time. Also, the visualization enables users to discover non-trivial information and patterns within events. To the best of our knowledge, this is the first tool that explicitly shows geopolitical links among locations given real-world events, allowing users to retrieve news by those relationships.

Chapter 6

Cartoglyphs: Visualizing Geographical and Geo-temporal Data with Glyphs

Several visualization techniques can be used to represent geographical data, such as maps, choropleths, cartograms, among others that may or may not preserve topology. Cartograms are thematic maps that scale a geographical region in proportion to some statistical value such as population, Internet use [90], scientific impact of publications [67], etc. This visualization technique dates back to 1870 [136] and is very popular for geo-referenced data. Cartograms can be categorized in 4 major types [180]: *contiguous*, *non-contiguous*, *Dorling* and *rectangular*. Contiguous cartograms deform the size of the regions of a map preserving topology, in particular adjacency. It is not possible to obtain perfect statistical accuracy and geographical accuracy at the same time, therefore algorithms that generated them must balance both features. Indeed, there are several algorithms that aim to generate this type of maps like the rubbermap cartograms by Tobler [175] or CartoDraw by Keim et al. [97]. Non-contiguous cartograms deform each region of the map independently, and therefore obtaining perfect statistical accuracy but not preserving the topology of the map. Dorling cartograms [52] use circles to represent each region, in which size is used to show a particular statistical variable. The larger the value of the variable, the larger the size of the circle. The position of each circle is computed in order to preserve its original geographic location in the map. However, this is not always possible and it is common that circles are moved to avoid overlapping. Finally, rectangular cartograms are similar to Dorling cartograms but use rectangles to represent each region. To evaluate which type of cartogram is best and in which contexts, Nusrat et al. [134] conducted a detailed study about the subject.

When time is included in the analysis, the problem of displaying geo-temporal data becomes much more complex, and designing effective visual representations can be very challenging. Two commonly used techniques used to address this problem are *small multiples* and *animation*. Small multiples, a term popularized by Edward Tufte [177], allows users to visualize several frames simultaneously using similar scale and visual encoding. This gives the user the advantage of viewing everything at once. This approach usually requires to reduce the visual representations in order to display them all in a screen with fixed size, which can be more difficult to analyze. Animation, on the other hand, represents each time frame as a full sized image, displaying each one of them after another, in a a sequence that conveys

the illusion of movement. Some studies indicate that an effective comparison between both techniques will depend on the task being carried out and exact experimental settings. For example, for observation of temporal changes in flow maps, Boyandin et al. [28] concluded that animation should be preferred for sudden change detection tasks, corroborating the study of Griffin et. al [76] on cluster detection. On the other hand, Robertson et al. [153] concluded that for trend detection in countries represented in a bubble chart, small multiples lead to more accurate results than animation. Nevertheless, both techniques can become too complex for geographical data, making it difficult for the user to compare quantity among locations, or recognize topology relationships among geographical entities. Regardless of the technique that is being used to analyze geo-temporal data, a reduced version of a representation of the world appears to be a reasonable option for reducing the complexity of the data. In this sense, one common technique used for visualizing multivariate data are glyphs. Glyphs include variables in a compact visual object, where each variable is mapped to a different visual channel.

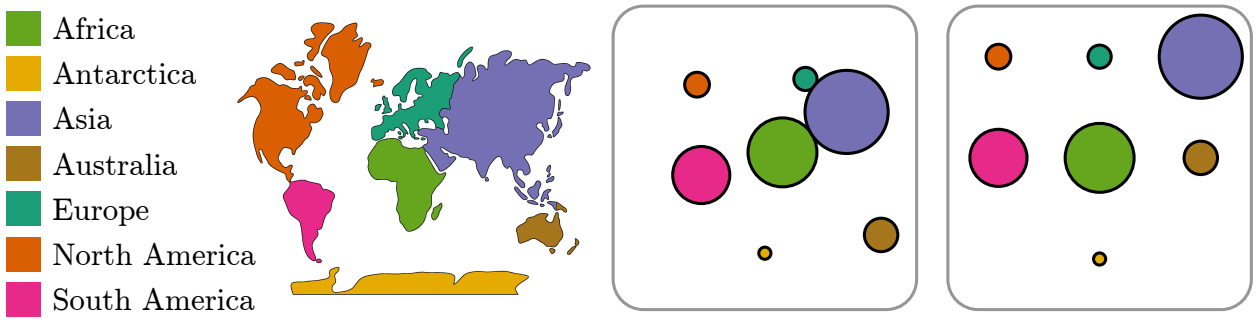
We propose to use simplified cartograms in the form of glyphs, which we call "Cartoglyphs". By using a simplification of Dorling and rectangular cartograms we have created simple glyphs that allow users to observe the evolution of geographical data over time. In this section, we present a preliminary design of Cartoglyphs and two case studies with data extracted from Twitter. As mentioned in chapter 3, despite of the use of glyphs over maps to visualize geographical variables, to the best of our knowledge our work is the first to use glyphs as the representation of the world itself.

6.1 Initial Designs

We present two initial version of Dorling cartoglyphs that represents the whole world. In this representation, we used continents as the administrative division level to display, following the convention that divides the world in seven areas: Asia, Africa, North America, South America, Antarctica, Europe, and Australia. We decide to use this convention as it gives us more information than, for example, those that consider America as one continent. We understand that this high level division can lead to information loss about the evolution of each individual country. Nonetheless, it allows us to have an overview and highlight interesting areas for further exploration.

An example of these two versions of cartoglyphs can be found in Figure 6.1. In this figure we use glyphs based on Dorling cartograms to represent the average population of the world. These cartoglyphs represents each continent as a circle colored with a distinctive color and which diameter represent the average population of each continent. We use two layouts to position each location in the glyph. The first is a centroid layout (Figure 6.1.b) in which each region is positioned in its centroid. For the continent administrative level displayed in Figure 6.1, we compute the centroid of each continent as the average of the centroids of each country that belongs to it, weighted by the area of the country. For the second layout, shown in Figure 6.1.c, all regions are arranged as a grid such that each region will be as closer as possible as its geographical location on a map. In other words, the glyph space is divided as a grid according to the number of locations to consider and then each location

is positioned in the center of the cell that is closer to its centroid in a map. This process is made automatically and it does not depend of the user interaction. All cartoglyphs were created with D3.js [45].



(a) Legend (b) Dorling with centroid layout (c) Dorling with grid layout

Figure 6.1: Dorling cartoglyphs representing average population by continent on 2010. Each circle represents a continent colored with a distinctive color which size is the average population per continent. (a) The legend of colors and geographical reference. (b) Dorling cartoglyphs with centroid layout, in which each region is positioned at the average continent centroid computed by considering the centroid of each country that compose it weighted by its area. (c) Represents a Dorling cartoglyphs with grid layout, where all continents are positioned in a grid such as each of them is closer as possible as its location in a map.

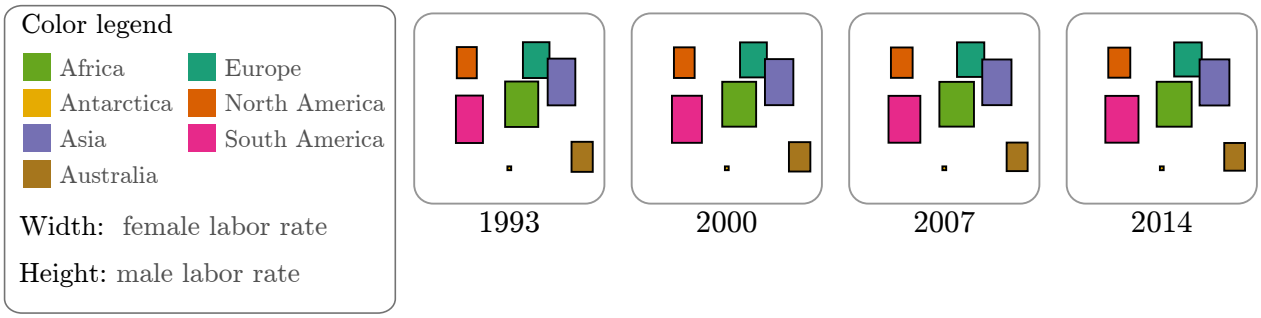


Figure 6.2: Preliminary rectangular cartoglyphs with centroid layout representing percentage of labor force by gender between 1993 and 2014, in four intervals. Each rectangle represents a continent colored with a distinctive color and positioned at the continent centroid. Their width is the percentage of female labor force and their height is the percentage of male labor force. Continents with more squared shape indicate a more equally distributed labor force by gender.

We also considered rectangular cartograms for building the glyphs, which can be seen in Figure 6.2. This figure shows four rectangular cartoglyphs representing male and female percentage of labor force over the world with data extracted from the World Bank [174]. With a similar intention that Polymetric views[107], this glyph use width and height to represent two different statistical values. In this case, the width of each rectangle represents the average percentage of female labor force rate per continent and the height is the average percentage of male labor force rate per continent. Like the previous figure, color is used to differentiate continents. With these four cartoglyphs is possible to observe than all continents tend to be squarified as the rate of women labor force increased over time. In addition, we can observe

that continents such as Europe and North America have a more equally distributed labor force among genders.

6.2 Cases Studies with Twitter Data

In this section we present two case studies in which cartoglyphs are used to analyze two news events detected from Twitter: the political impact on the Yemeni crisis and the missing Malaysia Airlines flight 370. For the first one we simply display them in sequence to analyze some dates, while for the second we arrange a set of glyphs in a tree layout to study two particular sides of the news event.

6.2.1 Observing Political Impact on Yemen Crisis

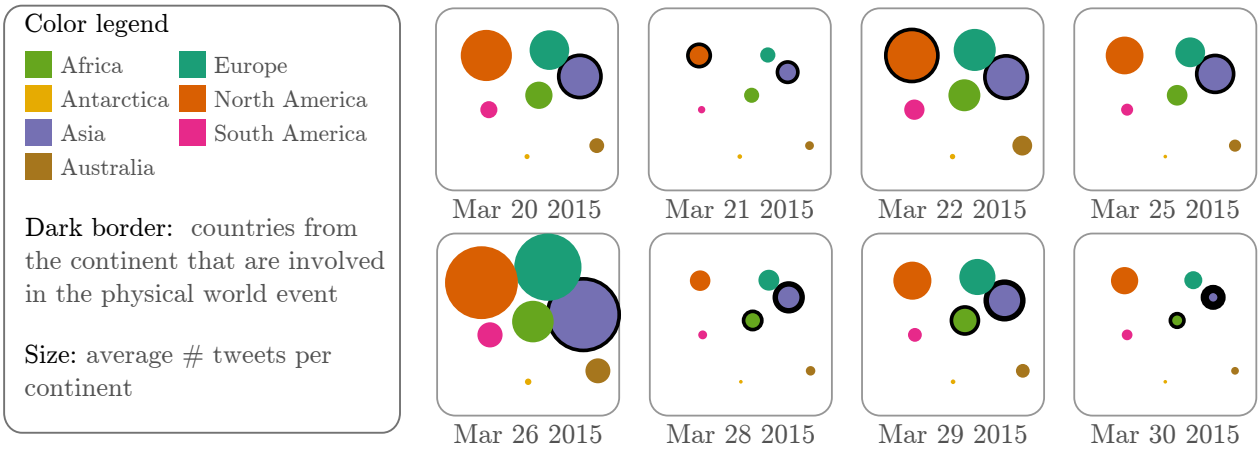


Figure 6.3: Propagation of tweets of a set of news events commenting about the Yemeni Civil War on March, 2015. Each selected news event is represented by a unique Dorling cartoglyph with centroid layout displaying seven continents, represented by a distinctive color. The size of each circle is the average number of tweets published from each continent. The border of the circle indicates the number of countries from that continent that are involved in the physical world news event.

We used a sequential set of Dorling CartoGlyphs to observe the evolution of geographical distribution of tweets commenting about the Yemeni Civil war that started in March, 2015 [198]. A selected subset news events represented with Dorling cartoglyphs are in Figure 6.3, made-up of eight non-contiguous time frames starting from 20th March, 2015, to 30th March, 2015. Using the same legend than Figure 6.1, each circle represents a continent. For this set of glyphs we used color hue to identify each continent and color opacity to indicate whether or not a country belonging to that continent is involved in the real-world event. In this case, as Yemen is the protagonist country of the event, Asia looks brighter in all frames, while other continents have a lower opacity value in most of the frames and only get "activated" in particular dates.

For example, in March 21th and 22th, North America and Asia are shown as activated. By inspecting the tweets and headlines on those dates, we observed that United States got involved in the conflict as US forces in Yemen evacuated an air base after al-Qaida seized a nearby city, on March 21th [119]. We also observe an increase in the participation in number of tweets between those dates, that could be explained by an increase of interest of the news.

We also observe that Africa also got involved in the event from March 28th to March 30th, that follows a similar pattern of increase of number of tweets between those days. By inspecting tweets on those days we realized Egypt and Saudi Arab got involved as the Arab League submit took place on those dates and Yemens President Hadi arrived in Egypt, on March 28th [8].

Finally, it is easy to observe that the event had the biggest impact on Twitter on March 26th, as the glyph on that date has the biggest circles. On this date, a Saudi air campaign was launched that resulted in the elimination of several Houthi leaders and big impact on Twitter [9]. In this example, we can observe the weakness of using a high level representation of the world, as it was not possible to observe that more than one country was involved. However, we expect to overcome this problem in future designs.

6.2.2 Following the Evolution of the Missing Malaysia Airlines Flight 370

We used a sequential set of Dorling cartoglyphs with centroid layout to observe the evolution of the geographical distribution of tweets commenting on the missing Malaysia Airlines flight 370, on March 2014 [197]. In this case, we used a tree layout to display the selected news to analyze this event.

In Figure 6.4 we observe a set of Dorling cartoglyphs with centroid layout representing a selected set of news detected from Twitter after the plane crashed. We show two subtopics of the news event, represented as two branches: (a) one concerning the investigation of the crash, and (b) one about the search for aircraft debris. Similar to Figure 6.1, in these Cartoglyphs each circle is a distinctively colored continent positioned on the centroid of the continent. The size of each circle represents the average number of tweets published from that continent about the news on that date. We use border width to indicate whether or not a country belonging to that continent was involved in the real-world event. In this case, as Malaysia is the main protagonist country of the event, Asia has a bold border in all frames, which makes it look "activated". On the other hand, other continents have no border, appearing "deactivated" in most frames. Regarding layout, we decided on a tree layout to show different subtopics as trees afford structured and systematic exploration [30].

When inspecting the investigation branch (a) of Figure 6.4, we observe two interesting events. The first one we observe that a country from North America got involved in the event on March 9, 2014, as its border is darker than the previous day. By inspecting tweets and headlines about the event, we saw that the United States sent FBI agents to help in the investigation on March 9th [170], and later there was news about the intentions of the FBI to analyze thumbprints of passengers who traveled with stolen passports [133]. This

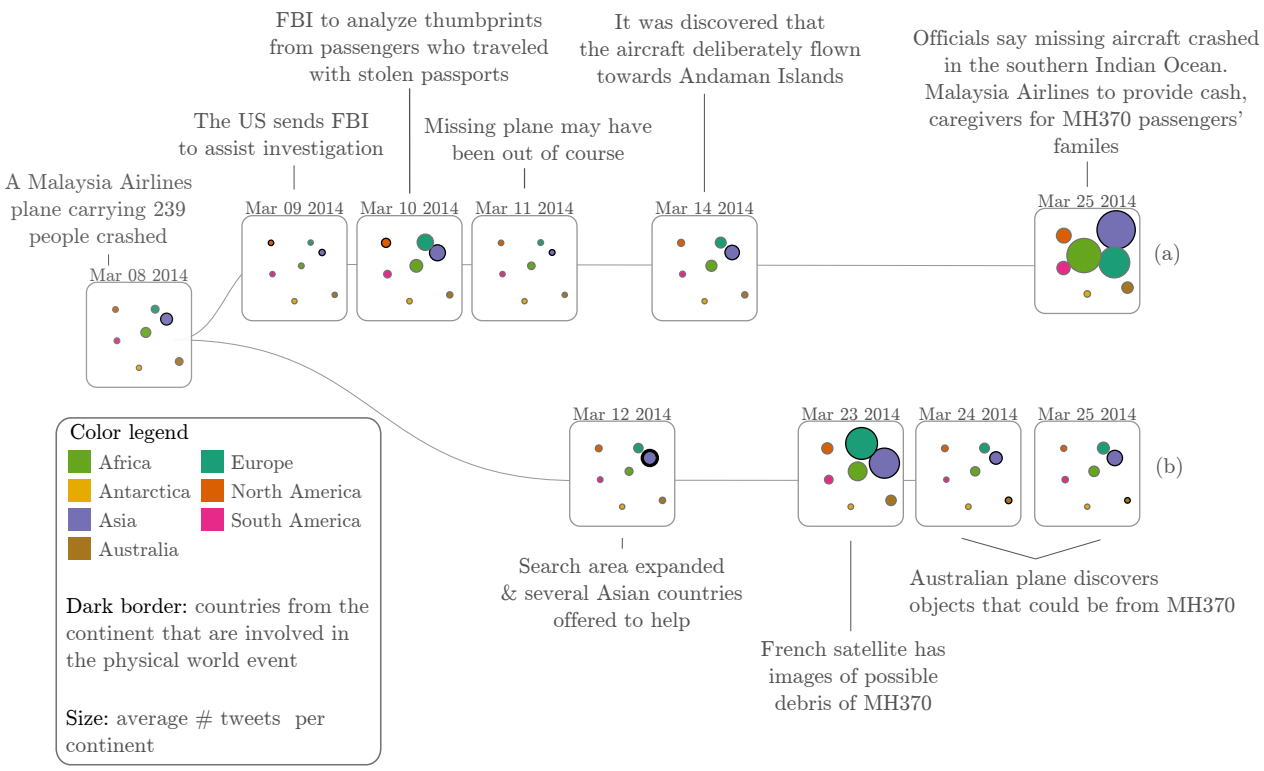


Figure 6.4: Propagation of tweets commenting about the missing Malaysia Airlines flight 370 on March, 2014. A set of selected news is represented by a unique Dorling cartoglyph with centroid layout displaying seven continents. The size of each circle is the average number of tweets published from each continent, which are represented by a distinctive color hue. Circles with a darker border represent continents with a country involved in the real world news event. The wider the border, the larger the number of countries involved. Two branches of the event are displayed: (a) the investigation of the news, and (b) the search from the plane debris.

country was the only one involved from North America that triggered that change. We can also observe that the impact of the news event increased on the next day as the circles on the glyph of March 10 have bigger size than those on March 9. This could indicate that United States is an influential country so people react to its involvement in this news. However, to determine the truth of this statement we would require another kind of analysis. By inspecting more the branch (a), we observe that the event with the biggest impact is located at the end of the timeline, which is the glyph for March 25, 2014. This news is about officials saying the missing aircraft crashed in the southern Indian Ocean, in addition to declarations of Malaysia Airlines about compensations to passengers' families.

On the other hand, in branch (b) of Figure 6.4, we observed three main states. The first one is on March 12th, when several Asian countries started to get involved in the search for debris as Asia has a bold border wider than on previous frames [124]. The second state is the "activation" of Europe, as this continent appeared with a bold border. By investigating this change, we discovered it was because Malaysia declared that France had satellite images of objects potentially from the missing plane [2]. This event had the bigger impact on twitter as the glyph that represent it has the circles with the bigger size. We can also observe how the

European interest on the subject grew on Twitter as the size of that continent is almost as big as Asia on that date. Finally, the two last frames on the timeline show that an Australian plane found some objects that could have been from Malaysia Airlines flight 370 [34].

6.3 Evaluation

In this section we present experimental results when comparing two Dorling cartoglyphs designs with a contiguous cartogram used ass glyph and a chart baseline. A summary of the visualizations to compare are in Figure 6.5. Our goal is to compare different variations of this type of representation and evaluate their effectiveness and efficiency for geographical and geo-temporal analysis. To reduce the number of variables to consider, we only focus on the static representation of a sequence of glyphs instead of animation and do not provide any type of interaction.

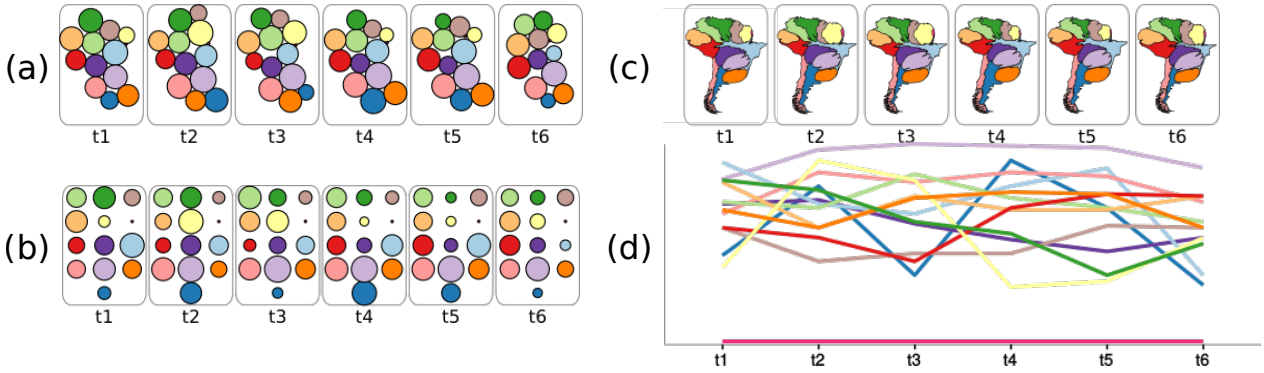


Figure 6.5: The four visualizations that were evaluated for the task of geo-temporal analysis: (a) Dorling cartoglyph with centroid layout, (b) Dorling cartoglyph with grid layout, (c) Contiguous cartoglyph, and (d) line chart. Each sequence represents the GDP value for each country starting from 1960 to 2010, with an interval of 10 years between each glyph.

Considering that our goal is to study the possibility of using glyphs to represent a piece of the world, we aim to explore if this simplification means that future users will not be able to recognize the locations inside each glyph. In addition, as the goal is to use this glyphs to analyze geographical and geo-temporal, we want to investigate if it is still possible to compare values associated to each location among the locations themselves and among different time windows. Therefore, we define our main research question as: *When using a glyph version of cartograms for geographical and geo-temporal analysis, which are the most important features for each task?* We divide this question into four specific ones:

- **RQ1:** How does the simplification of the shape of a region affect the performance of geographical data analysis? Does it follow the same behavior observed in other studies?
- **RQ2:** How does the abstraction of position and adjacency affect the performance of geographical data analysis?
- **RQ3:** How do these two variables (shape and position) affect in the analysis of geo-temporal data?

- **RQ4:** Does the number of locations represented affect the performance of geographical or geo-temporal data analysis?

We conducted a between subjects study to compare three cartogram based glyphs and a chart baseline. Subjects had to perform data analysis tasks derived from specific taxonomies.

The contribution of our work is to study the use of cartogram based glyphs for the analysis of geographical and geo-temporal data, in comparison to a chart based baseline. Our study focuses on three points aspects: (i) effectiveness, measured as the accuracy of answers, (ii) efficiency, measured as the time used to complete the tasks, and (iii) subjective metrics.

6.3.1 Evaluating Cartogram Related Visualizations

There are several studies of cartograms usability from a user perspective. Dent [48] evaluated contiguous cartograms and determined that people can achieve a better estimation of magnitude using contiguous cartograms when labels with value ranges are available. In particular, when testing user preferences, cartograms were confusing to read but interesting and innovative. He also concluded that users have better understanding of a contiguous cartogram if an inset map is provided. Griffin [77] reaches a similar conclusion of the benefits of a reference map when evaluating contiguous cartograms. Kaspar et al. [96] conducted an empirical study to compare contiguous cartograms with choropleths, combined with graduated circles for spatial inference making. They found that modified choropleths reported better overall performance but these results depended on the task complexity and the shape of the region that is being shown.

More specifically about users preferences, Sun and Li [166] conducted two experimental studies. In the first, cartograms were compared with thematic maps, and in the second, they compare different types of cartograms. They observe that pseudo-cartograms are more effective than Dorling cartograms both for quantitative data and quantitative data that included ordered classes. The effectiveness obtained for contiguous and non-contiguous cartograms were in the middle of these two maps. Han et al. [81] compared cartograms with the proportional symbol map for five tasks. They found that for comparing size at the country level, cartograms were less effective.

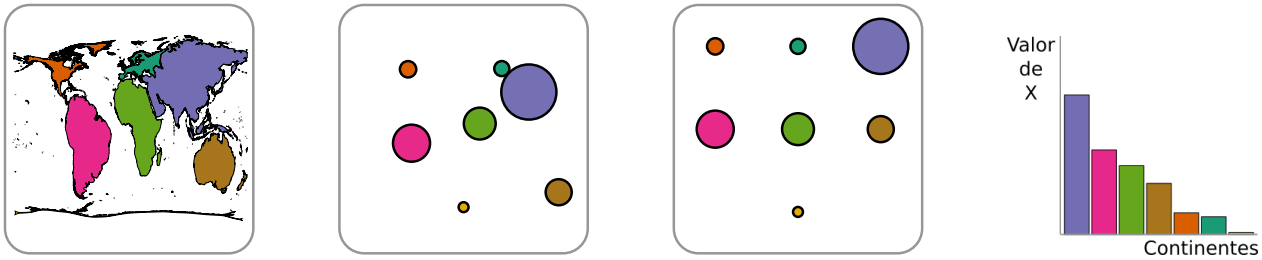
Nusrat et al. [134] provide one of the most complete cartogram evaluations that we were able to find in the literature. They compare the effectiveness of four main types of cartograms (contiguous, non-contiguous, Dorling and rectangular) for seven cartogram specific visualization tasks in the taxonomy of their early work [135]. They evaluated the time taken to complete each task, the error percentage of the answers and subjective metrics. They also provided a demographic analysis of performance by gender, age group, and educational level. They concluded that depending on the task, users performed differently with different types of cartograms. Our work is similar to thier in the sense that we conducted comparison between Dorling and contiguous cartogram for five of the seven cartogram-related tasks that they used.

Although there are several studies that evaluate glyph designs (e.g. Chernoff faces, Star

glyphs and spatial visualization [111]; impact of contours on the detection of data similarity with star glyph variations [62]; and for time series data [61]), ours is the first study that evaluates cartogram glyphs. In addition we extend existing evaluations in three aspects: First, we expand the list of analysis tasks by including specific ones for space in time. Second, we compare the effectiveness of cartograms with a visualization that is not specific to geographical data. Third, we study an alternative version of Dorling cartograms in which locations are organized as a grid.

6.3.2 Visualizations to Compare

We compare four visualizations. Three of these visualizations correspond to reduced or modified cartograms which are used as glyphs (e.g. Cartoglyphs). The other visualization is a chart, which we use as a baseline. We describe the as follows:



(a) (b) (c) (d)

Figure 6.6: The four visualizations to compare for the continent level. (a) Contiguous cartoglyph, (b) Dorling cartoglyph with centroid layout, (c) Dorling cartoglyph with grid layout, and (d) Bar chart.

- **Contiguous cartoglyph (Figure 6.6a):** we use a contiguous cartogram to represent an accurate version of the world in terms of topology. For this purpose we used the Gastner and Newman diffusion algorithm [68]. In other words, this is a traditional contiguous cartogram of reduced size.

- **Dorling cartoglyph with centroid layout (Figure 6.6b):** each geographical region is positioned in its centroid. For the continent administrative level, we compute its centroid as the average of all the centroids of each of its countries, weighted by their area. For the country administrative level, we use the capital's coordinates as the centroid.

- **Dorling cartoglyph with grid layout (Figure 6.6c):** similar to the previous visualization, the shape of each region is a circle with a given position. For this design, all regions are organized as a grid such that each region will be the closest possible to its geographical location on a map.

- **Charts:** This is a chart representation,which we used as baseline. We used a bar chart for displaying geographical data (Figure 6.6d) and a line chart for displaying geo-

temporal data (Figure 6.5d).

Since cartoglyphs were designed to be a simplification of the world, we do not expect them to include all the countries of the world in one glyph. For this reason, we argue that a cartoglyph using only continents will allow users to have an overview of the data and highlight interesting areas for further exploration. Nevertheless, we study cartoglyphs in two administrative levels: continents and countries, with the latest restricted to South America. The reason to do so is to study the effect on the performance of the participants when the number of locations is increased. For the glyph design at continent level, we followed the convention that divides the world in seven areas: Asia, Africa, North America, South America, Antarctica, Europe, and Australia. We decided to use this convention as it gives us more information than, for example, those that consider America as one continent. We understand that this high level division can lead to information loss about the evolution of each individual country. As explained earlier, we expect that this representation allows users to have high level understandinf of the data, which will hint areas to explore more in depth. For South America we considered thirteen countries: Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela. Figure 6.5 and Figure 6.6 show an example of each visualization for countries and continents, respectively.

Table 6.1: Dimensions in pixels of each visualization by type of location and type of analysis.

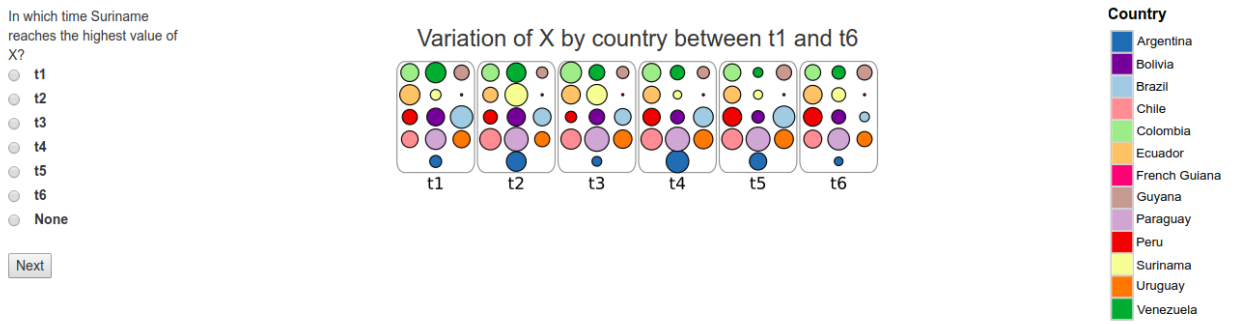
Visualization	Continents		South American countries	
	Geographical	Geo-temporal	Geographic	Geo-temporal
Contiguous cartoglyph	99x99	688x125	119x125	500x124
Dorling cartoglyphs with centroid layout	92x92	500x93	77x125	500x132
Dorling cartoglyphs with grid layout	92x92	500x94	87x124	499x133
Charts	331x288	500x187	278x257	499x186

Table 6.1 shows the size (in pixels) of each visualization according to location type and type of analysis. It is important to note that although we tried to set size as similar as possible among visualizations, this was not always possible. For example, for the contiguous cartoglyph of continents and geo-temporal analysis, we had to increase by 188 pixels the width of the image otherwise Europe was not recognizable. On the other hand, given that we were using bar charts are baseline, we did not apply size constrains and adjust a size that was enough to distinguish each bar.

6.3.3 Study Goal

Our main goal is to study which aspects should be considered when using a cartogram glyph. We do so by answering the questions proposed in Sec. 6.3. We focused on five tasks for the analysis of geographical data alone, and five tasks for geo-temporal analysis. We evaluated the accuracy of answers for effectiveness and time for efficiency, in addition to subjective metrics.

Figure 6.7: e is the question and the place to answer it, at the center is the visualization, and at the right there is the legend. The text of the image was translated from Spanish to English.



We work under four main hypotheses:

H1: For locations tasks, contiguous cartoglyphs should perform better than the two other cartoglpyhs as they preserve the topology of the map.

H2: For comparison tasks, in particular when comparing values over time, charts should report the best performance since they allow value comparison using only one dimension, in comparison with the other visualizations that require users to perform comparisons of areas.

H3: For comparison tasks, in particular when comparing values over time, Dorling cartoglyphs should report better performance than contiguous cartogyphs.

H4: For comparison and locations tasks, Dorling cartoglyphs with grid layout should have better performance than Dorling cartoglyphs with centroid layout as they have a fixed location, allowing users to find a location easier and compare them better.

6.3.4 Study Design

We designed a between subjects study in which each participant had to answer a set of questions using one of the four visualizations detailed in Sec. 6.3.2. We decided to exclude interaction as a variable in our experiment, therefore we used small multiples of glyphs to represent different time slices instead of animation. Given the number of questions presented, we decided to present only six time slices for the geo-temporal analysis to not overload participants.

Each participant had to go through two types of analysis: analysis of geographical data and analysis of geo-temporal data. We define the tasks for this two type of analysis from two taxonomies. The first one is from Nusrat et al. [135] that describe a set of tasks specifically for cartograms. The second is from Robert Roth [155], who define a set of tasks for geo-visualization categorized in three operand primitives: space-alone, attributes-in-space and space-in-time. We used the third operand for the geo-temporal tasks.

The tasks selected for each type of analysis are defined as follows:

Analysis of geographical data: conduct tasks oriented to cartographic tasks by analyzing one glyph at a time or a bar chart. For this type of analysis, users had to answer five questions, each related to a selected task defined as follows:

1. **Locate:** Nusrat et al. [135] define this task as finding the position of a region on a cartogram and Roth [155] defines it as finding in space when talking of the space-alone operand. Both definition are similar and focus on finding a particular location on a map. For the context of this study, this task is related to the action of finding a particular location on a glyph or chart. For this task we asked *"with which symbol is this location marked?"*, where the location was a particular continent or country.
2. **Compare:** this task asks for users to measure similar or different features among locations. This definition is used in other taxonomies in a similar way like Roth [155]. For this task we asked participants to answer if *"location A has a greater value than location B"*, where locations A and B would be either a continent or a country.
3. **Find greatest:** the objective of this task for the context of our study, is to find the location that has the greatest value of a particular statistic. A more general definition of this task is "Find top-k" [135], that include tasks like "find extremum" or "rank". Indeed, the taxonomy of Robert Roth [155] defines this task as "rank", which goal is to determine order or relative position among map features. For this task we asked *"which continent/country has the greatest value of the variable X?"*.
4. **Recognize:** For our study, the aim of this task is to recognize a particular location in the glyph. In the taxonomy by Nusrat et al. [135], this task's goal is to recognize the shape of a region from the original map. However, we will not focus on the shape of a location but on its position on a visualization and the possibility of recognize it. For this task we asked *"Which continent/country is marked with a cross?"*
5. **Find adjacency:** the goal of this task is to be able to find neighboring locations of a given region. This tasks is defined by Nusrat et al. [135] specifically for cartograms as some preserve topology while others do not. Since charts do not include any adjacency relationships, we exclude them for the analysis of this task.
   For this task we asked *"which location(s) is(are) geographically near to this particular location"*, where the particular location would be a specific continent or country. Given that this question had more than one possible answer, we decided to evaluate the accuracy of answers for this task using precision and recall by the formulas defined in Information Retrieval. With precision we mean the percentage of correct locations of all the locations answered by the participant and we compute it as $\frac{|participant\_answer \cap real\_answers|}{|participant\_answer|}$.
   We computed recall as $\frac{|participant\_answer \cap real\_answers|}{|real\_answer|}$ and is the percentage of the locations that were successfully identified.

Analysis of geo-temporal data: It consists of tasks oriented to the analysis of geographical attributes over time. For this type of analysis, we consider the five tasks defined by Roth [155] on the space-in-time operand primitive. The tasks are defined ad follows:

1. **Identify:** this task is defined as finding a specific value of a variable in time, or a spatial search with temporal constraint. For this task we asked *in which time a location reaches the greatest value of X?* Where a location could be a particular continent or country.

2. **Compare:** Similar to the definition for the analysis of geographical data, the objective of this task is to analyze the differences or similarities between geographical events including a temporal factor. It could also include the geographical patterns exhibited at different time periods, temporal composites, or temporal resolutions. We asked to participants *what can you say about the value of X for this location at these two specific times?*
3. **Rank:** the goal of this task is to order features in a map according to temporal proximity or other temporal characteristics. For example, ascending or descending. For this task, we asked: *which location(s) had a steady growth in the first three observations?*
4. **Associate:** This objective describes interactions tat characterize relationship among multiple map features. When applied for the space-in-time operand, it includes finding trends over time, or cause-effect relationships. We asked *what can you say about the tendency X for each location over time?*, where the location could refer to continents or countries depending to the stage the user is at. For these questions we give the participants the possibility to write their answers freely. To evaluate them, we coded their answers and assigned a score for each location. This score was 0 if the answers was incorrect or missing, 1 if it was partially correct, and 2 if it was correct. Then, the scores were averaged for all the locations. We normalized these answers and transformed them to a 0-100 scale.
5. **Delineate:** aims to structure the map features into logical components. This includes the division of data into distinct periods and finding peaks. For this task, we asked: *is there any location that has a peak? If that is the case, which ones and in which times?* Where location can refer to continents or countries depending to the stage of the study that the user is in. Similar to the task of *find adjacency* in the geographical analysis we computed answer precision and recall. In this case, the precision is the percentage of correct peaks of all the ones found by the participant. With recall we mean the percentage of the peaks that were successfully found by the participant.

It is important to note that the first three tasks of each type are very similar but with different focus. For most tasks, the possible answer was an option from a select or multi-select list, except for the tasks *associate* and *delineate* that required participants to write their answers in text.

When the participant arrived, the experimenter explained to her the goal of the experiment, the web interface and the type of questions they would have to answer. In particular, the experimenter presented the visualization used in the study and detailed each visual attribute until the participant said they understood. After that, the participant logged in and started with the first stage, in which they answered one question for each task for a visualization showing only continents. They later continued to a second stage, in which they did same but for South American countries. It is important to note that the location variable is a repeated measure. We provided a Web interface that recorded each participants answers, in addition to the time taken to complete each task. The interface had a simple design in which the questions were presented at the left, the visualization in the center, and the locations legend at the right (see Figure 6.7). We did not provide the users with a reference map for the cartograms, and only give users a color legend. The Web interface was displayed in screen with a resolution of 1280 x 1024 pixels.To measure subjective metrics we applied a post survey described in Appendix B and used the NASA TLX form [82].

To create the cartograms we used two datasets from The World Bank [174]. For creating the continent visualization we used the database of world population from 0 to 14 years and the GDP value for countries.

6.3.5 Participants

We recruited participants by posting announcements in common areas at the University of Chile. Out of the 60 people that participated in the study, 31 of them were men. Regarding their ages, 7 of them were less than 21 years old, 45 were between 21 and 30 years, and 8 were between 31 and 40 years. With respect to their educational background, 1 of them has PhD, 11 listed Masters, 24 listed undergrad, and 24 listed high school as their highest completed educational level. Regarding they familiarity with visualizations for data analysis, 10 reported not having experience, 18 reported having little experience, 19 reported having medium experience, and 13 reported having high level of expertise. None of the participants had extensive experience with cartograms. Finally, none of them declared to have any color vision impairment.

6.4 Results

6.4.1 Objective Metrics

In this section we report the results obtained for the objective metrics of both the analysis of geographical data and the analysis of geo-temporal data. For each task, we build a multilevel linear model that includes the location as a repeated measure, the visualization as a between subject variable and the interaction of both. In addition, we used contrasts to inspect significant differences further. We used the R statistical package version 3.4.1 to build the model, run the ANOVA and find differences with contrasts. A summary of the results for the geographical tasks are in Tables 6.2 and 6.3; those related to the geo-temporal tasks are in Tables 6.4 and 6.5.

Analysis of H1: Contiguous cartoglyphs did not outstand for all location tasks When analyzing the performance of contiguous cartoglyphs for location tasks, we did not find a significant difference for the *locate* and *recognize* tasks, neither for time nor for percentage error. For the analysis of the third location task, *finding adjacency*, we excluded charts for the analysis of those tasks because to complete them users depended on memory or previous knowledge, instead on the visualization itself. For this task, we did not find that contiguous cartoglyphs obtained better time performance than the other two, but participants using Dorling cartoglyphs with grid layout took statistically significant more time to complete the task than the other two visualizations. For the location precision, we observed there was a statistically significant difference between the visualization with $\chi^2(2) = 28.56, p < 0.0001$, again with Dorling cartoglyphs with grid layout obtaining statistically significant worse performance (with $p < 0.01$ for both visualizations). For location

Table 6.2: Time and % error for the four tasks for the analysis of geographic data. We report the ANOVA of the multilevel linear model build using location as a repeated measure and the visualization as a between subject variable, including the interaction of both variables.

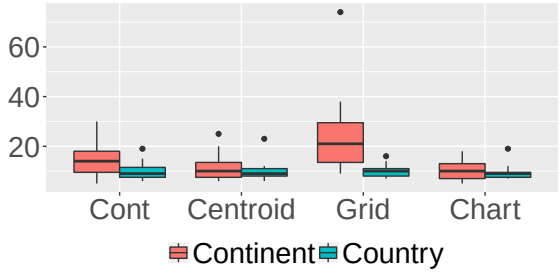
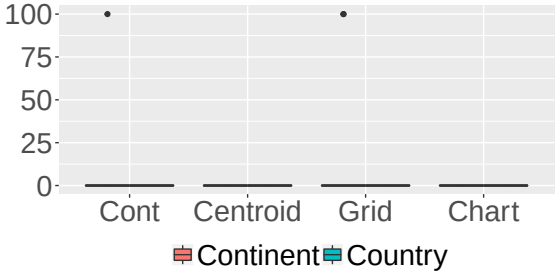
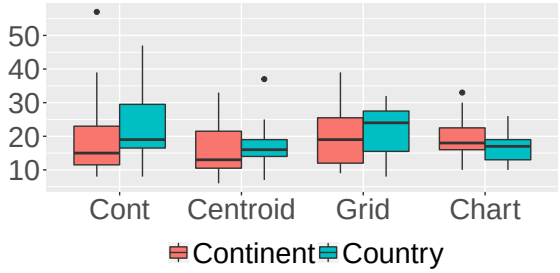
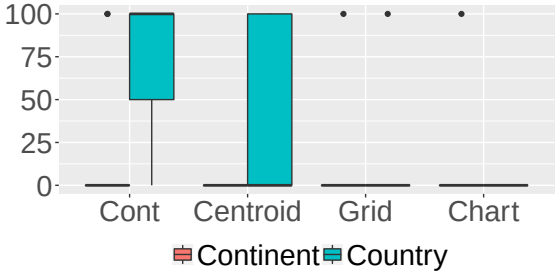
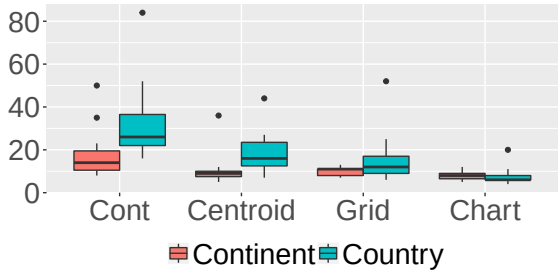
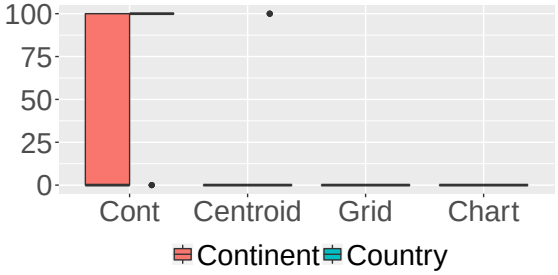
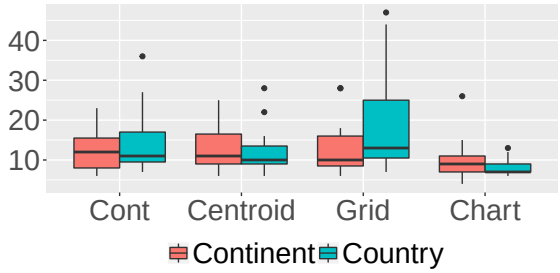
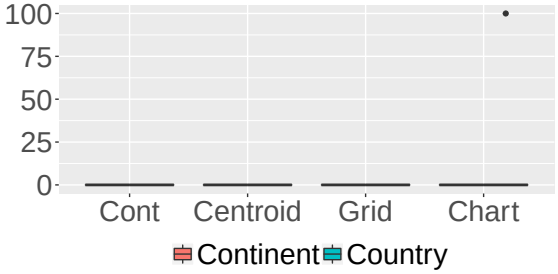
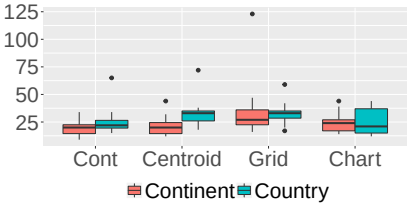
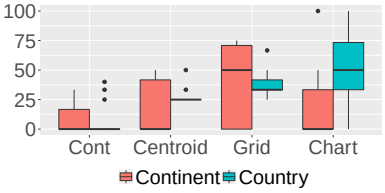
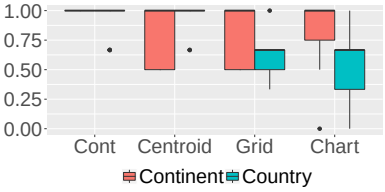
Task	Time (s)	Error%
Locate	location: $\chi^2(1) = 13.02, p = 0.0003$ *** vis: $\chi^2(3) = 15.91, p = 0.0012$ ** location x vis: $\chi^2(3) = 16.29, p = 0.0010$ ** 	location: $\chi^2(1) = 3.11, p = 0.078$ vis: $\chi^2(3) = 3.9, p = 0.27$ location x vis: $\chi^2(3) = 4.056, p = 0.26$ 
Compare	location: $\chi^2(1) = 0.36, p = 0.55$ vis: $\chi^2(3) = 4.97, p = 0.17$ location x vis: $\chi^2(3) = 2.68, p = 0.44$ 	location: $\chi^2(1) = 12.02, p = 0.0005$ *** vis: $\chi^2(3) = 23.47, p < 0.0001$ *** location x vis: $\chi^2(3) = 23.09, p < .0001$ *** 
Find greatest	location: $\chi^2(1) = 16.28, p = 0.0001$ *** vis: $\chi^2(3) = 32.98, p =< .0001$ *** location x vis: $\chi^2(3) = 12.66, p = 0.0054$ ** 	location: $\chi^2(1) = 7.85, p = 0.0051$ ** vis: $\chi^2(3) = 59.19, p =< .0001$ *** location x vis: $\chi^2(3) = 11.38, p = 0.0098$ ** 
Recognize	location: $\chi^2(1) = 0.59, p = 0.44$ vis: $\chi^2(3) = 10.29, p = 0.016$ * location x vis: $\chi^2(3) = 7.17, p = 0.067$ 	location: $\chi^2(1) = 1.01, p = 0.31$ vis: $\chi^2(3) = 3.09, p = 0.38$ location x vis: $\chi^2(3) = 3.17, p = 0.37$ 

Table 6.3: Time, location precision and location recall for the *Find Adjacency* task geographical data analysis. We report the ANOVA of the multilevel linear model build using location as a repeated measure and the visualization as a between subject variable, including the interaction of both variables.

Task	Time (s)	Location precision	Location recall
Find Adjacency	location: $\chi^2(1) = 2.45, p = 0.12$ vis: $\chi^2(2) = 8, p = 0.018$ * location x vis: $\chi^2(2) = 3.59, p = 0.17$ 	location: $\chi^2(1) = 0.18, p = 0.67$* vis: $\chi^2(2) = 28.56, p < 0.0001$ *** location x vis: $\chi^2(2) = 2.26, p = 0.32$ *** 	location: $\chi^2(1) = 0.12, p = 0.73$* vis: $\chi^2(2) = 26.38, p < 0.0001$ *** location x vis: $\chi^2(2) = 15.11, p = 0.0005$ *** 

recall, we did find that contiguous cartoglyphs obtained statistically significant better metrics for visualization with $\chi^2(2) = 26.38, p < 0.0001$, and for interaction visualization x location with $\chi^2(2) = 15.11, p < 0.001$. When inspecting this difference further we observed that contiguous cartoglyphs obtained better location recall than both Dorling cartoglyphs with $p < 0.01$.

To avoid bias in our data, we explored if there was any correlation between the participants' performance in the locations tasks and their declared knowledge about continents and countries. We computed the Pearson correlation coefficient and Spearman's rank correlation coefficient for the time taken to complete the tasks and the percentage error for their answers for the three locations tasks. We did not find any significant correlation between the task of the declared knowledge of the participant for any of the tasks.

With these results, we cannot confirm H1 in which we expected that contiguous cartoglyphs would obtain better results for all location tasks when compared with other visualization that not preserve accurate topology.

Analysis of H2: Charts outperform glyphs for certain tasks For this hypothesis we analyzed the geographical tasks *compare* and *find greatest*. Regarding the *compare* task, we did not find a significant effect for the time taken to complete it neither for location, nor for visualization, or the interaction between both. For the percentage error, we only found an statistically significant difference for the interaction between visualization and location with $\chi^2(3) = 23.09$ and $p < 0.0001$. For that interaction, we observed that charts had better performance than contiguous cartoglyphs ($p < 0.001$) and Dorling cartoglyphs with centroid layout ($p < 0.01$). Regarding the task of *finding the greatest*, we found that participants using charts had a statistically significant better performance in time ($p < 0.01$) and percentage error ($p < 0.001$) than contiguous cartoglyphs. When considering the interaction of visualization and location, there is also a significant better performance of charts over contiguous

Table 6.4: Time and error % for the four tasks for the analysis of geo-temporal data. We report the ANOVA of the multilevel linear model build using location as a repeated measure and the visualization as a between subject variable, including the interaction of both variables.

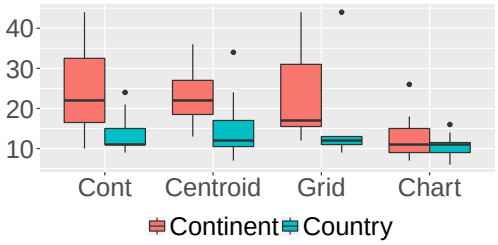
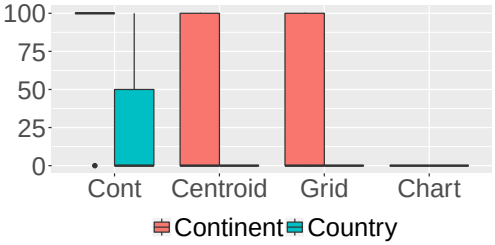
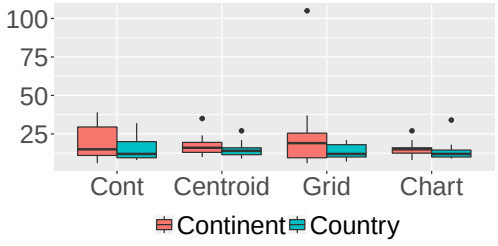
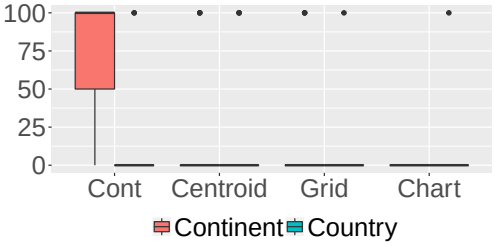
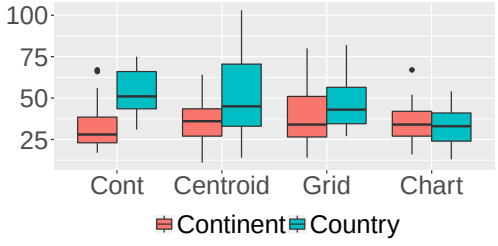
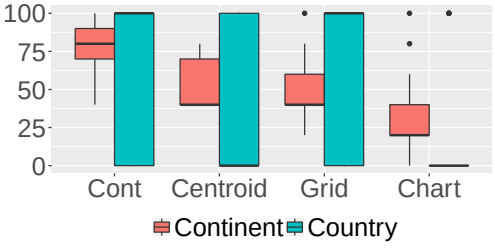
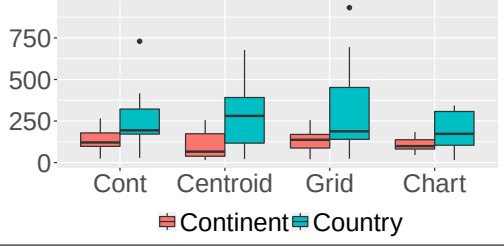
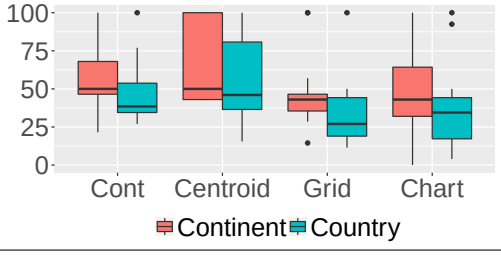
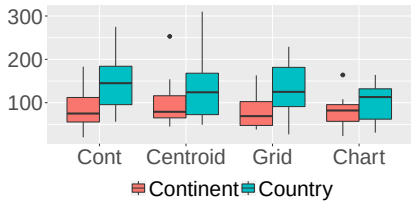
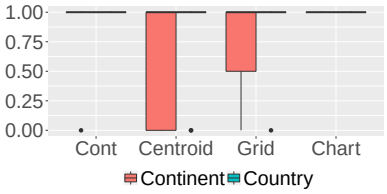
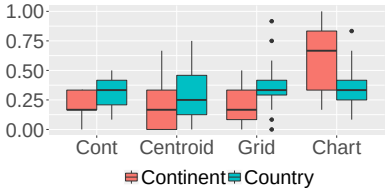
Task	Time (s)	Error%
Identify on time	location: $\chi^2(1) = 30.42, p < 0.0001$ *** vis: $\chi^2(3) = 16.47, p = 0.0009$ *** location x vis: $\chi^2(3) = 7.76, p = 0.05$ 	location: $\chi^2(1) = 27.41, p < 0.0001$ *** vis: $\chi^2(3) = 41.19, p < 0.0001$ *** location x vis: $\chi^2(3) = 18.12, p = 0.0004$*** 
Compare	location: $\chi^2(1) = 5.69, p = 0.017$ * vis: $\chi^2(3) = 2.56, p = 0.46$ location x vis: $\chi^2(3) = 3.34, p = 0.34$ 	location: $\chi^2(1) = 5.32, p = 0.021$ * vis: $\chi^2(3) = 18.31, p = 0.0004$ *** location x vis: $\chi^2(3) = 16.56, p = 0.0009$*** 
Rank	location: $\chi^2(1) = 10.36, p = 0.0013$ ** vis: $\chi^2(3) = 5.81, p = 0.12$ location x vis: $\chi^2(3) = 8.27, p = 0.04$* 	location: $\chi^2(1) = 3.01, p = 0.082$ vis: $\chi^2(3) = 16.86, p = 0.0008$ *** location x vis: $\chi^2(3) = 1.15, p = 0.76$ 
Associate	location: $\chi^2(1) = 32.52, p < 0.0001$ *** vis: $\chi^2(3) = 2.55, p = 0.47$ location x vis: $\chi^2(3) = 3.24, p = 0.36$ 	location: $\chi^2(1) = 12.96, p = 0.0003$ *** vis: $\chi^2(3) = 8.69, p = 0.034$ * location x vis: $\chi^2(3) = 0.039, p = 0.99$ 

Table 6.5: Time answer precision and answer recallfor the *Delineate* task for the analysis of geotemporal data. We report the ANOVA of the multilevel linear model build using location as a repeated measure and the visualization as a between subject variable, including the interaction of both variables.

Task	Time (s)	Answer precision	Answer recall
Delineate	location: $\chi^2(1) = 21.52, p < .0001$ *** vis: $\chi^2(3) = 4.17, p = 0.24$ location x vis: $\chi^2(3) = 3.29, p = 0.35$ 	location: $\chi^2(1) = 7.44, p = 0.0064$ ** vis: $\chi^2(3) = 8.70, p = 0.034$ * location x vis: $\chi^2(3) = 4.53, p = 0.21$ 	Answer recall location: $\chi^2(1) = 1.06, p = 0.30$ vis: $\chi^2(3) = 13.89, p = 0.0031$ ** location x vis: $\chi^2(3) = 21.26, p < 0.0001$ *** 

cartoglyphs with $p < 0.001$ for time and $p < 0.01$

Regarding the geo-temporal tasks, we observed that for the task of *identify in time* line charts had a better performance than the three glyphs for the time taken to complete the task ($p < 0.001$ for the three glyphs) and the percentage error (centroid: $p < 0.01$, grid: $p < 0.001$, contiguous: $p < 0.0001$) when considering the effect of the visualization alone. We also observed a significant difference when analyzing the interaction visualization x location for this task, for both the time (grid and contiguous: $p < 0.05$) and the percentage error (centroid: $p < 0.05$, grid: $p < 0.01$, contiguous: $p = 0.0001$). For the *compare* task we found that line charts had a statistically significant less percentage error than contiguous cartoglyphs with for the visualization alone ($p < 0.0001$), and the interaction visualization x location ($p = 0.001$). For the *rank* task, again participants using line charts had a significant lower percentage error than contiguous cartoglyphs, with $p < 0.01$ for the visualization alone. In addition, we observed a statistically significant increase of the time taken to complete the task when the number of locations increased when considering the visualization x location interaction for the contiguous cartoglyphs and Dorling with centroid layout, with $p < 0.05$ for both. Finally, for the *delineate* task we observed that participants using line charts had a significant lower percentage error for the recall of answers in comparison to the three glyphs, and for the precision of the answers for both Dorling cartoglyphs. Indeed, regarding the recall of answers, for the visualization alone we found $p < 0001$ for the three glyphs in contrast to the line charts, and for the visualization x location interaction we found a $p = 0.0001$ for contiguous cartoglyphs. For both Dorling cartoglyphs, we observed $p < 0.0001$ when analyzing the effect of the visualization alone in contrast to the line chart. For the precision of the answers of the *delineate* task, we observed that line charts had a better performance than both Dorling glyphs with $p < 0.01$ for the centroid layout and with $p < 0.05$ for the grid layout.

These results indicate that bar charts do not outperform all the three glyphs when being

used for comparison of geographical data alone: it was only better than contiguous cartoglyphs. On the other hand, when using the line charts, users reported better responses for several geo-temporal tasks, especially when being compared to the contiguous cartoglyphs.

Analysis of H3: Dorling cartoglyphs lead to better answers than contiguous cartoglyphs for some tasks of geo-temporal analysis For this hypothesis we observed that only for two of the geo-temporal oriented tasks, *identify in time* and *compare*, there was a clear difference between contiguous and both Dorling glyphs designs. Indeed, for the task of *identify in time* we observed that contiguous cartoglyphs had a statistically worse percentage error than both Dorling glyphs with $p < 0.001$ when considering the effect of the visualization alone. On the other hand, for the task of *compare* we observed that the effect of the visualization alone the value of $p = 0.0001$ for both Dorling glyphs when contrasting it with contiguous, with the latter having worst performance than the first two. For the same task, we observed that the number of locations also influenced the performance of the visualization as the visualization x location interaction was also statistically significant with $p < 0.01$ when contrasting contiguous with Dorlingwith centroid layout and $p < 0.05$ when contrasting contiguous with Dorling with grid layout.

These results partially confirms this hypothesis as two of the five geo-temporal tasks lead participants to have better answers with Dorling cartoglyphs than contiguous. Although, there were not significant different in the time taken to complete any of the tasks.

Analysis of H4: Centroid Layout vs Grid Layout for Dorling Cartoglyphs behave similarly For this hypothesis, we first inspected the tasks of geographical analysis and later those of geo-temporal analysis. For the geographical-oriented tasks, we observed that participants using the grid layout obtained a significant worse performance for two location tasks: *locate* and *find adjacency*. For the *locate* task, we observed that when analyzing the effect of the visualization alone, participants using the Dorling cartoglyphs with grid layout took more time to complete the task ($p < 0.0001$) and have a greater percentage error ($p < 0.05$) than those participants using the centroid layout. We also observed that the interaction between location and visualization had a significant effect between both glyphs, with $p < 0.01$. For the *find adjacency* task, we found that for the visualization alone, the grid layout was less efficient in time ($p < 0.01$) and less effective in the precision to identify locations ($p < 0.01$) than centroid layout. Similar to previous task, also the interaction between visualization and location had a significant effect with $p < 0.05$ for the time taken to complete the task.

Regarding the geo-temporal tasks, we did not find significant differences between both layouts except for the *associate* task. For this task, we found that participants with using the grid layout obtained a significant better performance than those using the centroid layout with $p < 0.05$.

With these results, we can not confirm the hypothesis that grid layout for Dorling cartoglyphs works better than the centroid layout. Indeed, for only two location oriented tasks the glyph with centroid layout showed worst performance. In addition, we did not find a

statistically significant difference for the geo-temporal tasks.

6.4.2 Subjective Metrics

In this section we discuss the results about the subjective metrics obtained by the NASA TLX form and the post survey. A summary of the results for the NASA TLX post survey is in Figure 6.8. We applied a Kruskal-Wallis test to analyze if there was difference among the visualizations for each of the factors of the NASA TLX factors. Although we did not find any statistically significant difference, we were able to observe some trends.

Regarding mental demand, our results suggests the following order from lower to higher: charts, Dorling with grid layout, Dorling with centroid layout, contiguous cartoglyph. When comparing the ones with best performance, charts and Dorling cartoglyphs with grid layout, and the one with he worst performance, contiguous cartoglyphs, we observed a non significant difference with $H(3) = 6.07$ and $p = 0.11$. A similar pattern was revealed for the frustration factor, with $H(3) = 4.29$ and $p = 0.23$.

This trend could be explained by the hypotheses **H2**, **H3** and **H4**, in which we expected that users would have better performance for comparison tasks for charts and Dorling cartoglyphs with grid layout. In particular, we received some positive comments about how participants used the invisible grid of the Dorling cartoglyphs with grid layout for make the comparison of circles easier.

Regarding physical demand ($H(3) = 1.63$, $p = 0.65$), we observe that the four visualization reported similar behavior. Nevertheless, similar to previous factors, both charts and Dorling cartoglyphs with grid layout obtained the lowest medians. We received some comments about how small were the maps and that sometimes it was hard for users to analyzed them. Indeed, a user commented that his eyes were hurting while conducting the tasks.

We were surprised by the results of performance and effort as we expected the tasks would not be difficult by participants. Indeed, we observed that participants thought they have an overall low performance ($M = 7.92$, $SD = 4.13$) and they have to make a great effort to achieve it ($M = 12.13$, $SD = 4.07$). We did not observe an statistically significant difference for both factors (performance: $H(3) = 0.49$, $p = 0.92$; effort: $H(3) = 2.35$, $p = 0.50$), however charts have the lowest median on both of them. For the last factor of the NASA TLX form, the temporal demand ($H(3) = 2.37$, $p = 0.50$) we did not observe significant difference among the visualizations.

Finally, for the post survey we asked three questions regarding their experience with the visualization:

- **How intuitive did you find the visualization?**: For this question we observed that charts had the higher score, followed by contiguous, Dorling with grid layout and Dorling with centroid layout. However we did not find any significant difference nor a significant trend.

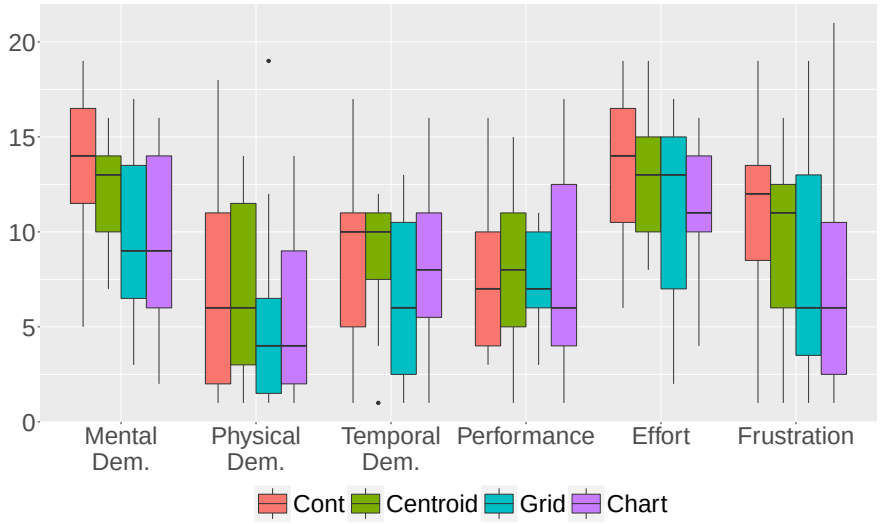


Figure 6.8: Summary of the 6 factors considered in the NASA TLX survey, for each of the for visualizations compared in this study.

- **How difficult was to compare the values of X across the time slices?**: For this question we observed that contiguous cartoglyphs received a a higher score of perceived difficulty for analysis than the other three visualizations, however we did not find a statistically significant difference among them. In particular, when comparing grid and centroid layout for Dorling cartoglyphs, we observed a trend in which participants found centroid layout easier for values comparison. This result is interesting because when inspecting the differences between centroid and grid layout for Dorling cartoglyphs in **H4**, the only significant difference among them was for the *associate* task in which the grid layout obtained better performance than the centroid layout. Therefore, although participants found that with centroid layout it was easier to compare values among time slices, the associate task shows different.

- **How difficult was to find each region of the visualization?**: Although we did not observed a statistically significant difference among the four visualizations for this question ($H(2) = 6.97$, $p = 0.072$), we observed a trend in which contiguous cartoglyphs obtained the lower difficulty, followed by Dorling with grid layout, charts and finally Dorling with centroid layout. The preference of users for the design of Dorling cartoglyph is contradictory with the results found in **H4** in which we compared both Dorling designs were compare in terms of performance. We aim to inspect this difference in further research.

6.5 Discussion

In this section we discuss our results and propose some recommendation for design based on them.

6.5.1 Analysis of Research Questions

We discuss our findings for each research question proposed in previous sections.

RQ1: How does the simplification of the shape of a region impact in the performance of geographical data analysis? Does it follow the same behavior observed in other studies?

For this question we compared the contiguous cartoglyph with both Dorling designs as regions are represented with different shapes with those designs. We observed a significant difference for only two geographical data analysis tasks: *find greatest* and location recall for *find adjacency*. For the first task we observed that participants using contiguous were more prone to do errors that participants using any of the Dorling designs. For the second task, we observed that participants were more likely to identify a greater percentage of correct adjacent locations with contiguous than with both Dorling cartoglyphs design. With these results, we observe than although shape impacted in the performance of participants, it was only significant for some of the selected tasks.

When comparing our results with those reported by Nusrat et al. [134], we found some considerable differences. For example, for the *locate* task their work found a statistically significant difference between contiguous cartogram and Dorling Cartogram for the percentage error, which we did not observed in our results. Another difference is regarding the *find greatest* in which, although both work found statistically significant difference between contiguous and Dorling for the percentage error, the distribution of values is not the same. Indeed, while in the work of Nusrat participants using contiguous cartograms show better performance than Dorling, our results show the opposite. These differences could be due several reasons like size of stimulus, number of locations displayed, type of locations analyzed, or question proposed for each task, among others. We aim to explore the cause of differences in future work.

RQ2: How the abstraction of position and adjacency impact in the performance of geographical data analysis?

To answer this question we compared the behavior of both Dorling cartoglyph designs. As detailed earlier, for the tasks *locate* and *find adjacency*, participants using Dorling with grid layout presented worst performance than those using Dorling with centroid layout. In this aspect, the position of each location impacted in the analysis of tasks that were related to geographical location. We did not found any significant difference for comparison tasks between both Dorling layouts. We found this result surprising as we expected that by fixing each location to a position in the glyph, it would be easier for users to compare values and identify each locations. We did find that participants used this feature as we received comments about it, however this did not impacted significantly in their performance.

RQ3: How these two variables impact in the analysis of geo-temporal data?

We observed that for two geo-temporal data analysis tasks both Dorling cartogyphs design had statistically significant less percentage error than contiguous cartoglyphs (*identify in time* and *compare*). On the other hand, for the task *delineate* we observed that contiguous had a

significant worse perform than Dorling with centroid layout. These results indicates that the shape used to represent each region impacted in the performance for geo-temporal analysis. On the other hand, when comparing the behavior of both Dorling cartoglyph to understand of the position impact in the performance of the selected geo-temporal tasks we observed that only for one task (*associate*) there was a significant difference, which indicate us that there is not much impact in the performance of participants when considering the change of the position of each location for the selected geo-temporal tasks.

RQ4: Does the number of locations impact on the performance of geographical or geo-temporal data analysis?

This question is based on the expectation that given an increase in the number of locations being analyzed would impact on the performance of participants in a negative way. In other words, the larger the number of locations, the larger time they would require to answer and they would be more prone to error. Given the collected data, we observed that the number of locations impacted on the performance in time and accuracy with statistical significance for several of the tasks analyzed. However, we were surprised to find that this fact did not impacted in all the tasks as we initially expected.

6.5.2 Recommendations for Design

Given the results of our study, we propose the following recommendations for the use of cartoglyphs and charts.

For the analysis of geographical data:

- If the adjacency of the locations being analyzed is critical, contiguous cartoglyphs could lead to better results than other designs.
- If the task requires to find greatest values, Dorling cartoglyphs could lead to more accurate and efficient results than contiguous.
- When choosing between centroid layout and grid layout for Dorling cartoglyphs, centroid layout should be chosen over grid layout as the later could lead to more error for some tasks that require identifying geographical characteristics of the data.

For the analysis of geo-temporal data:

- If the comparison of values between time slides is critical, charts should be preferred over cartoglyphs.
- There is no significant difference between centroid and grid layout, leaving the decision to the user preference. Grid layout could be useful for some of them as they use the imaginary grid as a guide for comparison.

In particular, we propose that Dorling cartoglyphs provide a good balance between displaying information about the topology of a map and still allowing comparison between

values.

6.5.3 Open Questions

Given our initial results, we have several open question that will be addressed in future research:

- **How the results variate when the size of the stimulus variate?** We identified that there were significant changes among the visualizations studied, however we have not considered the size of the stimulus as a variable to analyze. Which is the recommended size for each glyph? How does it variate when there are several glyphs to observe at the same time? Does it exist a limit in which the size of each glyph is not distinguishable?

- Similar to previous question, **to which extend can we simplify the world so participants can still distinguish each locations?** Does this depend of the number of considered locations?

- **Is it possible to not use color as a variable to represent each region?** Given that our experiment design used color to represent each region, in future work we want to investigate if users can remember each location without this visual variable and rely in position or memory.

- **Is it possible to include more variables such as border size or color in order to make them multivariate?** For this study we only included one statistical variable at each time frame, does cartoglyphs support more variables at the same time?

- **Which interactions could improve the type of data analysis we are studying?** Is it better to use animation? Maybe brush and linking to relate the same locations?

6.6 Considerations to Validity

There are several considerations to validity in our study which are presented in this section. The first consideration is low diversity of the representation of each task. Indeed, given that each task was represented by only one question in our study, we understand that the tasks may be not represented enough. Second, we only considered a limited set of locations to be analyzed so we don't have information about how the selected designs can be generalized to other regions such as Europe or the United States. However, it is important to note that the goal of cartoglyphs is to display a simplified version of the world. Therefore it might not be efficient to use them to display all the countries of the world or all the cities of a country at once. In the same line, the selected continents and countries used in the present study are real regions which participants have seen before even if they are not completely familiar with their the exact location. Regarding this issue, we could have used locations that are less familiar to users like postal codes or electoral districts to avoid preconceptions about

spatial processes like the work of Beecham et al. [22]. Nevertheless we consider important to study cartoglyphs in a familiar scenario like continents and countries of South America as they represent possible visualization participants could see in their geopolitical context. Third, given that questions and locations were not randomized there could have been a learning effect that could impact in our results. We aim to analyze this factor in future work and compare those results with those presented here. Finally, an important limitation when studying the performance of visualization designs for geo-temporal analysis is the number of time frames considered. We only considered six time slices with the goal of no overload participants, which could be a low number when dealing with real data. Nevertheless, we found our work an important starting point to consider a map as glyph, instead of glyphs over maps, for the analysis of geographical and geo-temporal data.

6.7 Summary

Designing visualizations for the analysis of geo-temporal data can still be a challenge. In this chapter we presented preliminary designs of cartoglyphs, a simple visual representation of the world. This representation aims to allow users to conduct geo-temporal analysis in more effective ways.

We first showed how Dorling and rectangular cartograms can be reduced to a glyph to decrease the complexity of the visualization of geographical data. In addition, we presented two initial layouts to display the geographical locations displayed in the glyph. We applied Dorling Cartoglyphs to show the evolution of geographical distribution of tweets commenting on two news events over several days. In addition, by using them we detected when new countries got involved in the real-world event and how these changes impacted in the reaction of Twitter users.

We later studed the proposed visualization to evaluate its effectiveness and efficiency. More specifically, we conducted a study with 60 people to analyze their accuracy of answers and the time taken to complete a series of tasks. Five of those tasks were only for geographical data and five were for geo-temporal data, both derived from tasks taxonomies from the literature. We compared four visualizations: three cartogram based glyphs and charts in two administrative levels of locations, continents and countries.

We found that the number of locations can influence the performance of the analysis for some tasks but not for all of them. We also concluded that although two glyphs maps could use the same visual variable for quantity, the position of the elements being compared can influence the answers of participants when analyzing data with them. Also, we hypothesized that charts would have a better performance in comparison to the glyphs but that was not true for several tasks.

Chapter 7

Conclusion and Perspectives

In this section we summarize the contributions of this dissertation and discuss possible directions for future work.

7.1 Contributions and Conclusions of the Dissertation

The contributions of this dissertation and their corresponding conclusion, are divided into three main components:

1. **Geo-temporal Representation of Events Extracted from Social Media**: We presented a high level event representation that characterizes the geo-temporal components of an event extracted from social media. This representation not only focuses on the locations where an event happened, but also it considers the relevant geopolitical entities involved and the places to where a news event was propagated (Chapter 4). More formally, we define two types of locations. The first, *protagonists locations*, corresponds to those geopolitical entities involved in a real world event. The interaction among these types of locations denote conflicts, alliances or neutral relationships that represent historical past events that can allow us to understand the present. The second type of location is defined as *interested locations*, which are the places from where people commented on an event in social media.

   With these location type definitions we describe two kinds of scopes: (i) *provenance*, which indicates if an event is local, regional or global in terms of the protagonists locations involved; and (ii) *impact*, which denotes whether an event is local, regional or global depending on the number of locations that were interested in commenting on the event.

   The formal definition of this representation allows researchers to compute event similarity based on protagonists or interested locations. In addition, it enables to compute similarity metrics among locations based on events in which they are protagonists together or showed similar interest on social media.

We then evaluated this event representation with an exploratory analysis using a two-year database of news events extracted from Twitter. We observed that this representation allowed us to explore international relations among countries and how they evolved over time.

With this first contribution, we addressed the first main hypothesis and its sub-hypotheses presented in the introduction.

H1: *The data published in social media platforms contains valuable information about what is happening in the real-world.*

- ***H1.1****: Analyzing data from social media yields historical data about news related to geopolitical interaction among countries as consequence of news events.*
- ***H1.2****: By analyzing social media data one can understand how people reacted to a news event and the geographical places to which news events propagated to.*
- ***H1.3****: By analyzing data from social media one can obtain insight of how events relate to each other over time.*

The data analysis conducted in Chapter 4 allowed us to extract valuable information about news events that happened in the physical wold. In particular for ***H1.1***, the geopolitical analysis reported in that chapter confirmed the historical value of that data. For example, we observed a high event similarity between Ukraine and Russia as protagonists locations. This similarity corresponds to the long-term conflict between both nations, which was reflected in more events being extracted where they were protagonists. Regarding ***H1.2***, we observed that some countries showed similar level of interest on events that linked particular countries. This is the case of interested locations that commented on events that Brazil and Germany were protagonists. This phenomenon reflected the events related to the FIFA World Cup of 2014. Finally, for ***H1.3*** we saw the similarity among countries evolved over time. For instance, the previously mentioned event about the Crimean crisis describes a pattern in which both nations display a high similarity metric for the period that the crisis lasted. On the other hand, sporadic events such as the disappearance of the Malaysia Airlines flight MH370 on March 2014, describe a sporadic spike for a shorter period of time.

2. **Galean**: The second contribution of this dissertation is the interactive visual interface to explore news events in the proposed event representation presented in Chapter 5. The interface of this visualization tool display events over a map by their protagonists countries and easily understand their provenance. In addition, it allows users to inspect an event in more details to observe the distribution of interested locations and the impact of the event. Finally, users can read the messages related to an event for more detailed information about it. The messages are categorized by their source: they can come from news Twitter accounts or regular users.

   To validate the usefulness of the tool we conducted two case studies and two users studies. In the first case study we followed the provenance of the events related to the Crimean crisis. By observing the frequency and category of events, we concluded how

the crisis started being a local event and evolved to be of regional and global provenance. The event developed mainly at regional scope given that Ukraine and Russia were the main protagonists of the events. However, the United States also intervened at some point so several events were of global provenance. The second case study was about the Nepal earthquake of 2015. The motivation of this case study was to use Galean to retrieve past events that allowed users to understand a recent one. More specifically, given a later event about a financial loan from Japan to Nepal for its recovery we used the tool to search for related news events that allowed us to understand it.

We conducted two user studies for the tool evaluation. The first one was a qualitative study with expert users. In this evaluation we aimed to obtain feedback about whether Galean could be used for daily work by journalists. We concluded that Galean could not necessarily be used by journalists in their daily work as it will depend on the tasks they need to carry out. Also in this evaluation we obtained important usability feedback that allowed us to improve the tool. The second evaluation that we conducted was a quantitative study in which we compared Galean to a competitive baseline in terms of objective and subjective metrics. In terms of objective metrics we measured the efficiency and effectiveness for the task of retrieving information about relationships between countries. For these metrics, Galean was more efficient than the baseline for the tasks proposed this task. For subjective metrics, participants who used Galean showed less frustration when conducting the tasks and were more confident with the information displayed than what was in the baseline. These results offer us evidence on the usefulness of Galean in comparison to a competitive baseline.

In this chapter we addressed **H2.1** and **H2.2**:

H2*: Visual representation of news in their geopolitical context allows users to extract valuable knowledge about the real world.*

- ***H2.1****: An expressive visual representation allows users to visually identify and extract patterns, which cannot be easily found through manual or quantitative analysis of the raw data.*
- ***H2.2****: An expressive visualization of the geopolitical context of a news event allows users to extract relationships among events and participating entities.*

Indeed, regarding ***H2.1***, given the two users studies conducted with Galean, we confirm that users were able to extract geopolitical relationships of news events from its interface. In addition, the designed interface proved to be efficient in terms of the time taken to extract this information and to provide less frustrating ways of displaying it. From the qualitative study in this chapter, we can confirm ***H2.2***, as expert users were able to identify real world events and follow the evolution of the protagonists locations that participated in them.

3. **Cartoglyphs**: The last part of the dissertation presents and describes the evaluation of Cartoglyphs, cartogram based glyphs. Their goal is to provide a simplified version of the world in order to help users conduct geo-temporal analysis. This last part is presented in Chapter 6. We first describe the initial designs for Cartoglyphs, using

both rectangular and Dorling cartograms. We later describe two case studies in which we use them to follow the evolution of news events extracted from Twitter. In the first one, we followed the propagation of tweets commenting on the Yemeni Civil War between the 20th and 30th of March, 2015. In this case we positioned the glyphs in a grid layout. Among the patterns we observed, we saw that the day with the biggest impact on Twitter described the news about the launch of a Saudi air campaign in which several Houthi leaders were eliminated. In the second case study we used a tree layout to display two branches of the development of the news regarding the missing Malaysia Airlines flight 370. The first branch was about the investigation of the news and the second about the search of the plane debris. In this case study we observed several patterns of people commenting on the news events, allowing us to find out which the countries were involved in the real world event.

Later in the chapter we describe the formal evaluation of Cartoglyphs that we conducted. We tested two layouts to position locations inside the glyph when using Dorling cartogram, a contiguous cartoglyphs design, and charts as baseline. The goal of he study was to study if the simplification of shape and position to represent a location impacts the performance of geographical data analysis, among other research questions. We concluded that these features impacted in the performance of users,depending on the task carried out.

With this contribution we started to explore hypothesis **H2.3**:

H2.3*: A simple visual representation of geographical data allows users to extract knowledge from several points of view of a news event.*

Although we used them to represent two geographical variables of news events extracted from Twitter, we still need to formally study if they are effective for the analysis of more than one variable of geographical data. The idea of Cartoglyphs is new and the study we conducted gave us the first insights to arrive at an effective design for a simplified version of the world in order to representing multiple variables at the same time.

7.2 Future Work

The future work is divided in three components:

1. Further analyze the model and data we already have. For example, we would like to explore not only the relationships among countries but also understand their influence. In this sense, we would like to answer questions like: *Did the impact of the event change after a particular country got involved?*.

   On the other hand, we would like to enhance the news event representation. Some of the features we would like to explore are:

   - Increase the granularity on the time associated to an event, incorporating not only the date when it was detected, but also its duration and when it finish. This will

allow us to more finely analyze the speed that a news event was propagated or how it becomes of local, regional o global impact, for example.

- Conduct sentiment analysis of the content of tweets to determine if the link between two countries is a conflict or an alliance between them. In particular, we would like to study the strength of the alliances or conflicts found. This type of analysis can be applied to the case of the Crimea crisis, to study the evolution of the tension among the countries involved.

- Include entities to study other kinds of influences. For example, we would like to answer questions such as, *does a tweet by a particular famous person influence the way news gets propagated?*

2. Improve Galean from two perspectives: (i) identify the real need of final users or change the focus of final users, and (ii) improve its interface. Regarding the first one, we would like to conduct field studies to identify which cases this tool would most benefit the work of journalists. In this sense, we have considered the possibility of contacting users from other research areas. In particular, we have had informal conversations with political scientists who declare that this kind of analysis could be useful for their daily work. We aim to explore the possibility of collaborating with them in order to improve Galean and to help them in their research. Regarding the improvement of Galean's interface, we would like to consider the following topics:

   - Improve the way Galean displays tweets commenting on a news event by organizing the message via subtopics or a visual approach like ThemeRiver [85].

   - Include a visual representation of the evolution of a news event. This could be done by using Cartoglyphs to represent a news event, either in a linear or tree layout. We expect to include this visual representation once the research of Cartoglyphs is more advanced.

   - Include visual representations of others aspects of a news event such as sentiment analysis or a network of the influential people commenting on the event.

3. Continue the exploration of Cartoglyphs for visualizing multivariable geographical data. In future work we expect to deepen the analysis of subjective metrics and how they relate to the objective ones. We also will compare how demographics of participants could influence in their performance, and if there is any interaction of those variables. Additionally, we will study the gaze movement when participants conducted this analysis for the different visualizations, type of locations and tasks. To gather more qualitative feedback of cartoglyphs, we will to conduct focus groups with possible final users such as geographers or sociologists. Furthermore, we would like to address the questions proposed in Chapter 6:

   - How does the effectiveness of Cartoglyphs variate when their size is changed?

   - To what extent can we simplify the world representation of Cartoglyphs so participants can still distinguish each location? Does this depend on the number of

locations being considered?

- Is it possible to not use color as a variable to represent each region so users identify a location only by their position inside the glyph?
- Is it possible to include more visual variables such as border size of a location in order to make them multivariate?
- Which interactions could improve the type of data analysis we are studying?

Finally, when the theoretical aspects of the visual features of Cartoglyphs are further studied, we want to include them in an interactive visual interface in which each glyph is positioned according to a certain layout, allowing users to discover patterns on geographical data. More specifically, we would like to explore:

- which layouts are effective to position glyphs in order to find similar behavior of geographical data,
- how these layouts can also include time in order to preserve temporal linearity.

Appendix A

Related Surveys and Questionnaire for User Study Conducted to Evaluate Galean (Chapter 5)

A.1 Questionnaire for news event analysis

The list of questions asked for the news analysis are the following:

- In which date the news event started?
- List the countries that were involved to the news event (separated by a comma)
- When X country got involved in the news event for the first time? We asked for Saudi Arabia as X for the news about the Yemen rebels, and China for the news about the Malaysia Airline MH360 lost.
- Give between 5 and 10 relevant keywords about the event (separated by a comma)
- How much impact do you the event had in X date? We asked for March 26th, 2015 as X for the news about the Yemen rebels, and March 8th, 2014 for the news about the Malaysia Airline MH360 lost.

A.2 Pre-survey: Demographic Information

The pre-survey for the user study we conducted to evaluate Galean is originally in Spanish. Next are the questions translated in English:

- Gender:

  - Woman
  - Men
  - Prefer not to disclose
- Age:
  - Less than 21 years old
  - Between 21 and 30 years old
  - Between 31 and 40 years old
  - Older than 40 years old.
- Greatest educational level achieved:
  - Undergraduate
  - Master
  - PhD
- Greatest educational level on process:
  - Undergraduate
  - Master
  - PhD
- How familiar are with the use of visualization for data exploration?
  - None: I have never used a visualization for data exploration.
  - Low: I almost never use visualizations for data exploration.
  - Medium: I regularly use visualizations for data exploration.
  - High: I use visualizations for data exploration very frequently and/or it is part of my job.
- How frequently do you read international news?
  - I never read about international news
  - Very few times, only when something important happened.
  - regularly, I read international news once or twice a week

- Very frequently, I read everyday about international news
- Which is your English level?
  - None
  - Low
  - Medium
  - High

A.3 Post-survey: Galean Interface

The post-survey conducted to evaluate the subjective perception of Galean by the participants. The original survey was conducted in Spanish and next we enumerate the questions translated in English.

- How intuitive did you find the interface? Answer as a 5 points liker scale with 1 being "Not intuitive" to 5 being "very intuitive".
- Would you use it to analyze news events? Answer as a 5 points liker scale with 1 being "I would never use it again" to 5 being "I would use it frequently".
- How confident were you in the information displayed? Answer as a 5 points liker scale with 1 being "very low confidence" to 5 being "very high confidence".
- Did you lose notion of time while conducting the task? Answer as a 5 points liker scale with 1 being "not much" to 5 being "a lot".
- Would you recommend the tool? Answer as a 5 points liker scale with 1 being "I would not recommend it" to 5 being "I would certainly recommend it".
- How much satisfied are you with the tool? Answer as a 5 points liker scale with 1 being "not much satisfied" to 5 being "very satisfied".
- How much information do you think the interface did not allowed you to see? Answer as a 5 points liker scale with 1 being "few" to 5 being "a lot".
- Any extra comment about the interface?

Appendix B

Post-survey for Cartoglyphs Study (Chapter 6)

The post survey for the Cartoglyphs study was divided in two main sections: demographic information and subjective perception of the inspected glyph.

Demographic information:

- Gender:
  - Woman
  - Men
  - Prefer not to disclose
- Age:
  - Less than 21 years old
  - Between 21 and 30 years old
  - Between 31 and 40 years old
  - Older than 40 years old.
- Greatest educational level achieved:
  - Undergraduate
  - Master
  - PhD
  - Other: specify

- Greatest educational level on process:
  - Undergraduate
  - Master
  - PhD
  - Other: specify
- How familiar are with the use of visualization for data exploration?
  - None: I have never used a visualization for data exploration.
  - Low: I almost never use visualizations for data exploration.
  - Medium: I regularly use visualizations for data exploration.
  - High: I use visualizations for data exploration very frequently.
- How easy is for you to recognize the world continents? Answer as a 5 points liker scale with 1 being "very difficult" to 5 being "very easy".
- How easy is for you to recognize the countries of South America? Answer as a 5 points liker scale with 1 being "very difficult" to 5 being "very easy".
- In which data related area do you work?
  - Computer science
  - Journalism
  - Sociology
  - Other: specify

Subjective perception of the visualization:

- How intuitive did you find the visualization? Answer as a 5 points liker scale with 1 being "Not intuitive" to 5 being "very intuitive".
- How difficult was for you to compare the values of X between each time slice? Answer as a 5 points liker scale with 1 being "Not difficult" to 5 being "very intuitive".
- How difficult was for you to find each region inside the glyph? Answer as a 5 points liker scale with 1 being "Not difficult" to 5 being "very intuitive".
- Any comment?

Bibliography

[1] Gapminder. https://www.gapminder.org/. Accessed: 2016-06-22.

[2] ABC News. Malaysia Airlines MH370: French satellites spot objects in missing jet search area. http://www.abc.net.au/news/2014-03-23/missing-malaysia-airlines-plane-mh370-france-satellite-images/5339788, 2014. Accessed March 4, 2018.

[3] H. Abdelhaq, C. Sengstock, and M. Gertz. EvenTweet: Online localized event detection from Twitter. *Proceedings of VLDB Endowment*, 6(12):1326–1329, Aug. 2013.

[4] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proc. VLDB Endow.*, 6(12):1326–1329, Aug. 2013.

[5] Abdur Chowdhury. Global pulse. https://blog.twitter.com/official/en_us/a/2011/global-pulse.html, 2011. Accessed on May 27, 2018.

[6] A. Agarwal, R. Singh, and D. Toshniwal. Geospatial sentiment analysis using twitter data for uk-eu referendum. *Journal of Information and Optimization Sciences*, 39(1):303–317, 2018.

[7] C. C. Aggarwal and K. Subbian. Event detection in social streams. In *SDM*, page 624635, 2012.

[8] Al Arabiya English. Egypt calls for joint Arab military force at summit. http://english.alarabiya.net/en/News/middle-east/2015/03/28/Yemeni-president-arrived-in-Egypt-for-Arab-League-summit-.html, 2018. Accessed on January 18, 2018.

[9] Al Arabiya English. Saudi Decisive Storm waged to save Yemen. http://english.alarabiya.net/en/News/middle-east/2015/03/26/GCC-states-to-repel-Houthi-aggression-in-Yemen-statement-.html, 2018. Accessed on January 18, 2018.

[10] Alexa Internet. https://www.alexa.com/siteinfo/twitter.com, 2018. Accessed April 10, 2018.

[11] Aljazeera. Ukraine drops EU plans and looks to Russia. http://www.aljazeera.com/news/europe/2013/11/

ukraine-drops-eu-plans-looks-russia-20131121145417227621.html, 2013. Accessed November 14, 2017.

[12] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. 1998.

[13] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, page 3745, 1998.

[14] O. Alonso, S. Bannur, K. Khandelwal, and S. Kalyanaraman. The World Conversation: Web Page Metadata Generation from Social Sources. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 385–395, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[15] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

[16] D. Archambault, D. Greene, P. Cunningham, and N. Hurley. Themecrowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 77–84. ACM, 2011.

[17] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10, Jan 2011.

[18] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

[19] BBC News. Peru-Chile border defined by UN court at The Hague. http://www.bbc.co.uk/news/world-europe-25911867, 2014. Accessed on November 3, 2017.

[20] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11(2011):438–441, 2011.

[21] J. Beddow. Shape coding of multidimensional data on a microcomputer display. In *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, pages 238–246, 478, Oct 1990.

[22] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood. Map lineups: Effects of spatial structure on graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):391–400, Jan 2017.

[23] A. Bergel. Roassal. http://agilevisualization.com/, 2018. Accessed September 16, 2018.

[24] Berico Technologies. CLAVIN: Cartographic Location And Vicinity INdexer. http:

//clavin.bericotechnologies.com/, 2012–2017. Accessed November 14, 2017.

[25] K. Blandford, R. DAmour, K. Leasor, A. Terry, and I. V. de Latour. Citizen security and social media in mexicos public sphere. *VIRTUalis*, 4(7):13–40, 2014.

[26] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (STARs)*, pages 39–63, 2013.

[27] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, Dec. 2013.

[28] I. Boyandin, E. Bertini, and D. Lalanne. A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. *Computer Graphics Forum*, 31(3pt2):1005–1014, 2012.

[29] A. Brambilla, R. Carnecky, R. Peikert, I. Viola, and H. Hauser. Illustrative flow visualization: State of the art, trends and challenges. *Visibility-oriented Visualization Design for Flow Illustration*, 2012.

[30] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE transactions on visualization and computer graphics*, 20(12):2271–2280, 2014.

[31] D. Buscaldi and P. Rosso. A conceptual densitybased approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313, 2008.

[32] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2649–2658, Dec 2012.

[33] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[34] CBS News. Malaysia Airlines Flight 370 search: Chinese, Australian planes spot "objects" in Indian Ocean. https://www.cbsnews.com/news/malaysia-airlines-flight-370-search-chinese-plane-spots-suspicious-objects-in-in 2014. Accessed March 4, 2018.

[35] T. Chadefaux. Early warning signals for war in the news. *Journal of Peace Research*, 51(1):5–18, 2014.

[36] J. Chae, D. Thom, Y. Jang, S. Kim, T. Ertl, and D. S. Ebert. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, 38:51–60, 2014.

[37] D. Chakrabarti and K. Punera. Event Summarization Using Tweets. In *International AAAI Conference on Web and Social Media*, 2011.

[38] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[39] J. Choi and W. B. Croft. Temporal models for microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2491–2494, New York, NY, USA, 2012. ACM.

[40] D. H. Chung, P. A. Legg, M. L. Parry, R. Bown, I. W. Griffiths, R. S. Laramee, and M. Chen. Glyph sorting: Interactive visualization for multi-dimensional data. *Information Visualization*, 14(1):76–90, 2015.

[41] CNN. Saudi-led coalition strikes rebels in Yemen, inflaming tensions in region. `http://edition.cnn.com/2015/03/26/middleeast/yemen-saudi-arabia-airstrikes/`, 2015. Accessed November 14, 2017.

[42] M. Cordeiro and J. Gama. *Online Social Networks Event Detection: A Survey*, pages 1–41. Springer International Publishing, Cham, 2016.

[43] P. Craig, N. R. Seïler, and A. D. O. Cervantes. Animated geo-temporal clusters for exploratory search in event data document collections. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 157–163. IEEE, 2014.

[44] A. Croitoru, A. Crooks, J. Radzikowski, and A. Stefanidis. Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, 27(12):2483–2508, 2013.

[45] D3.js. `https://d3js.org/`, 2015. Accessed 22 November 2016.

[46] DBpedia. `http://dbpedia.org`, 2017. Accessed on November 3, 2017.

[47] B. De Longueville, R. S. Smith, and G. Luraschi. "OMG, from Here, I Can See the Flames!": A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 73–80, New York, NY, USA, 2009. ACM.

[48] B. D. Dent. Communication aspects of value-by-area cartograms. *The American Cartographer*, 2(2):154–168, 1975.

[49] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122, Oct 2010.

[50] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405, 2015.

[51] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, Nov. 2008.

[52] D. Dorling. Area cartograms: their use and creation. In *Concepts and techniques in modern geography*. Citeseer, 1996.

[53] W. Dou, K. Wang, W. Ribarsky, and M. Zhou. Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, pages 971–980, 2012.

[54] M. Drk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, Nov 2010.

[55] Elastic. ElasticSearch. `https://www.elastic.co/products/elasticsearch`, 2018. Accessed September 24, 2018.

[56] N. Eltantawy and J. B. Wiest. The arab spring— social media in the egyptian revolution: reconsidering resource mobilization theory. *International Journal of Communication*, 5:18, 2011.

[57] T. Ertl, J. Chae, R. Maciejewski, H. Bosch, D. Thom, Y. Jang, and D. S. Ebert. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 143–152, Washington, DC, USA, 2012. IEEE Computer Society.

[58] Event Registry. Event registry system. `http://eventregistry.org/`, 2015. Accessed November 23, 2017.

[59] Fox 2 News. Teenager shot, killed in Ferguson apartment complex. `http://fox2now.com/2014/08/09/man-shot-killed-in-ferguson-apartment-complex/`, 2014. Accessed November 14, 2017.

[60] Fox News. President Obama removing trade benefits for Russia over Ukraine. `http://www.foxnews.com/politics/2014/05/07/president-obama-removing-trade-benefits-for-russia-over-ukraine.html`, 2014. Accessed November 14, 2017.

[61] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3237–3246. ACM, 2013.

[62] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini. The influence of contour on similarity perception of star glyphs. *IEEE transactions on visualization and computer graphics*, 20(12):2251–2260, 2014.

[63] R. Fuchs and H. Hauser. Visualization of multi-variate scientific data. In *Computer Graphics Forum*, volume 28, pages 1670–1690. Wiley Online Library, 2009.

[64] Fdration Internationale de Football Association (FIFA). Navas-inspired Ticos win shootout, reach quarters. `http://www.fifa.com/worldcup/matches/round=255951/match=300186459/match-report.html`, 2014. Accessed on November 9, 2017.

[65] K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[66] J. Galbraith, I. Moustaki, D. J. Bartholomew, and F. Steele. *The analysis and interpretation of multivariate data for social scientists*. CRC Press, 2002.

[67] S. Gao, Y. Hu, K. Janowicz, and G. McKenzie. A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 204–213, New York, NY, USA, 2013. ACM.

[68] M. T. Gastner and M. E. Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7499–7504, 2004.

[69] P. Gatalsky, N. V. Andrienko, and G. L. Andrienko. Interactive analysis of event data using space-time cube. *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 145–152, 2004.

[70] GDELT. The news co-ocurrence globe. `http://data.gdeltproject.org/blog/news-cooccurrence-globe/globe_cooccur.html`, 2013-2014. accessed 23 August 2017.

[71] Global Voices. `http://globalvoicesonline.org/`, 2016. Accessed 22 November 2016.

[72] U. GmbH. GeoNames. `http://geonames.org/`, 2017. Accessed November 14, 2017.

[73] A. Godavarthy and Y. Fang. Cross-language microblog retrieval using latent semantic modeling. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 303–306, New York, NY, USA, 2016. ACM.

[74] Google Inc. Google's Geocoding API. `https://developers.google.com/maps/documentation/geocoding/intro`, 2005. Accessed November 14, 2017.

[75] M. J. Greenacre. *Biplots in practice*. Fundacion BBVA, 2010.

[76] A. L. Griffin, A. M. MacEachren, F. Hardisty, E. Steiner, and B. Li. A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. *Annals of the Association of American Geographers*, 96(4):740–753, 2006.

[77] T. Griffin. Recognition of areal units on topological cartograms. *The American Car-*

tographer, 10(1):17–29, 1983.

[78] A. Guille and C. Favre. Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach. *CoRR*, abs/1505.05657, 2015.

[79] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[80] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

[81] R. Han, Z. Li, P. Ti, and Z. Xu. Experimental evaluation of the usability of cartogram for representation of globeland30 data. *ISPRS International Journal of Geo-Information*, 6(6), 2017.

[82] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139 – 183. North-Holland, 1988.

[83] M. Hasan, M. A. Orgun, and R. Schwitter. A survey on real-time event detection from the twitter data stream. *Journal of Information Science*, page 0165551517698564, 2017.

[84] S. Hassan, J. Sanger, and G. Pernul. SoDA: Dynamic visual analytics of big social data. In *Big Data and Smart Computing (BIGCOMP), 2014 International Conference on*, pages 183–188, Jan 2014.

[85] S. Havre, B. Hetzler, and L. Nowell. Themeriver: visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 115–123, 2000.

[86] D. Henry, E. Stattner, and M. Collard. Information propagation routes between countries in social media. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1295–1298, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[87] S. R. Hong, Y.-S. Kim, J.-C. Yoon, and C. R. Aragon. Traffigram: Distortion for clarification via isochronal cartography. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 907–916, 2014.

[88] T. Huet, J. Biega, and F. M. Suchanek. Mining history with Le Monde. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 49–54, New York, NY, USA, 2013. ACM.

[89] A. Jadhav, H. Purohit, P. Kapanipathi, P. Anantharam, A. H. Ranabahu, V. Nguyen, P. N. Mendes, A. G. Smith, M. Cooney, and A. Sheth. Twitris 2.0: Semantically empowered system for understanding perceptions from social data. In *Semantic Web*

Challenge, International Semantic Web Conference (ISWC), 2010.

[90] T. Johnson, C. Acedo, S. Kobourov, and S. Nusrat. Analyzing the evolution of the internet. In *17th IEEE Eurographics Conference on Visualization (EuroVis–short papers)*, volume 17, 2015.

[91] J. Kalyanam, M. Quezada, B. Poblete, and G. Lanckriet. Prediction and characterization of high-activity events in social media triggered by real-world news. *PLOS ONE*, 11(12):1–13, 12 2016.

[92] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 667–678, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[93] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[94] M. Karimzadeh, W. Huang, S. Banerjee, J. O. Wallgrün, F. Hardisty, S. Pezanowski, P. Mitra, and A. M. MacEachren. GeoTxt: A web api to leverage place references in text. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 72–73, New York, NY, USA, 2013. ACM.

[95] A. Karve and M. Gleicher. Glyph-based overviews of large datasets in structural bioinformatics. In *Information Visualization-Supplements, 2007. IV 2007. 11th International Conference on*, pages 1–6. IEEE, 2007.

[96] S. Kaspar, S. Fabrikant, and P. Freckmann. Empirical study of cartograms. In *25th International Cartographic Conference*, volume 3, page 5, 2011.

[97] D. A. Keim, S. C. North, and C. Panse. Cartodraw: A fast algorithm for generating contiguous cartograms. *IEEE Transactions on Visualization and Computer Graphics*, 10(1):95–110, 2004.

[98] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. Ebert. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017.

[99] A. Kitamoto and T. Sagara. Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '12, pages 23–26, New York, NY, USA, 2012. ACM.

[100] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[101] C. Korson. Political agency and citizen journalism: Twitter as a tool of evaluation. *The Professional Geographer*, 67(3):364–373, 2015.

[102] M.-J. Kraak. The space-time cube revisited from a geovisualization perspective.

[103] T. Kraft, D. X. Wang, J. Delawder, W. Dou, Y. Li, and W. Ribarsky. Less after-the-fact: Investigative visual analysis of events from streaming twitter. In *Large-Scale Data Analysis and Visualization (LDAV), 2013 IEEE Symposium on*, pages 95–103. IEEE, 2013.

[104] J. Krumm and E. Horvitz. Eyewitness: Identifying local events via space-time signals in twitter feeds. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '15, pages 20:1–20:10, New York, NY, USA, 2015. ACM.

[105] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[106] H. Y. Kwon and Y. O. Kang. Risk analysis and visualization for detecting signs of flood disaster in twitter. *Spatial Information Research*, 24(2):127–139, Apr 2016.

[107] M. Lanza and S. Ducasse. Polymetric views-a lightweight visual approach to reverse engineering. *IEEE Transactions on Software Engineering*, 29(9):782–795, 2003.

[108] H. W. Lauw, E.-P. Lim, H. Pang, and T.-T. Tan. Stevent: Spatio-temporal event model for social network discovery. *ACM Trans. Inf. Syst.*, 28(3):15:1–15:32, July 2010.

[109] Leaflet. `http://leafletjs.com/`, 2015. Accessed November, 13 2017.

[110] C.-H. Lee, H.-C. Yang, T.-F. Chien, and W.-S. Wen. A novel approach for event detection by mining spatio-temporal information on microblogs. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 254–259, July 2011.

[111] M. D. Lee, R. E. Reilly, and M. E. Butavicius. An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24*, pages 1–10. Australian Computer Society, Inc., 2003.

[112] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pages 1–10. ACM, 2010.

[113] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9), 2011.

[114] P. A. Legg, D. H. Chung, M. L. Parry, M. W. Jones, R. Long, I. W. Griffiths, and M. Chen. Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. In *Computer Graphics Forum*, volume 31, pages 1255–1264. Wiley Online Library, 2012.

[115] A. E. Lie, J. Kehrer, and H. Hauser. Critical design and realization aspects of glyph-based 3d data visualization. In *Proceedings of the 25th Spring Conference on Computer Graphics*, pages 19–26. ACM, 2009.

[116] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei. Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):2451–2466, Nov 2016.

[117] X. Liu, Q. Li, A. Nourbakhsh, R. Fang, M. Thomas, K. Anderson, R. Kociuba, M. Vedder, S. Pomerville, R. Wudali, R. Martin, J. Duprey, A. Vachher, W. Keenan, and S. Shah. Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 207–216, New York, NY, USA, 2016. ACM.

[118] P. Longley. *Geographical information systems and science*. Wiley, Chichester, 2005.

[119] Los Angeles Time. Yemen officials: U.S. forces evacuate as Al Qaeda takes city. `http://beta.latimes.com/world/middleeast/la-fg-yemen-us-forces-20150321-story.html`, 2018. Accessed on January 18, 2018.

[120] L. R. Lucchesi and C. K. Wikle. Visualizing uncertainty in areal data with bivariate choropleth maps, map pixelation and glyph rotation. *Stat*, 6(1):292–302, 2017. sta4.150.

[121] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: GeoTwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181–190, Oct 2011.

[122] J. Maldonado, V. Peña Araya, and B. Poblete. Spatio and temporal characterization of chilean news events in social media. TAIA '15, 2015.

[123] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.

[124] Mashable. Oil Rig Worker: 'I Saw the Malaysia Airlines Plane Come Down'. `https://mashable.com/2014/03/12/malaysia-airlines-370-search-area/#jYblZL8yQZqL`, 2014. Accessed March 4, 2018.

[125] J. Mellon and C. Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[126] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. Van Harmelen. Semantic technologies for historical research: A survey. *Semantic Web*, 6(6):539–564, 2014.

[127] D. Metzler, C. Cai, and E. Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 646–655, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[128] J. Meyer-Spradow, L. Stegger, C. Döring, T. Ropinski, and K. Hinrichs. Glyph-based spect visualization for the diagnosis of coronary artery disease. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1499–1506, 2008.

[129] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[130] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11(5th):25, 2011.

[131] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[132] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010.

[133] NBC News. Who Flew? FBI to Check Thumbprints of Impostor Passengers. `https://www.nbcnews.com/storyline/missing-jet/who-flew-fbi-check-thumbprints-impostor-passengers-n49156`, 2014. Accessed March 4, 2018.

[134] S. Nusrat, M. J. Alam, and S. Kobourov. Evaluating cartogram effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016.

[135] S. Nusrat and S. Kobourov. Visualizing cartograms: Goals and task taxonomy. *arXiv preprint arXiv:1502.07792*, 2015.

[136] S. Nusrat and S. Kobourov. The state of the art in cartograms. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: State of the Art Reports*, EuroVis '16, pages 619–642, Goslar Germany, Germany, 2016. Eurographics Association.

[137] S. Oeltze, A. Hennemuth, S. Glaßer, C. Kühnel, and B. Preim. Glyph-based visualization of myocardial perfusion data and enhancement with contractility and viability information. In *VCBM*, pages 11–20, 2008.

[138] Oxford Dictionaries. http://www.oxforddictionaries.com/definition/english/event?q=event, 2017. Accessed on October 24, 2017.

[139] J. H. Parmelee. The agenda-building function of political tweets. *New Media & Society*, 16(3):434–450, 2014.

[140] J. Pearlman and P. Rheingans. Visualizing network security events using compound

glyphs from a service-oriented perspective. In *VizSEC 2007*, pages 131–146. Springer, 2008.

[141] V. Peña-Araya, M. Quezada, B. Poblete, and D. Parra. Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using twitter. *EPJ Data Science*, 6(1):25, Oct 2017.

[142] S. Peters and L. Meng. Visual analysis for nowcasting of multidimensional lightning data. *ISPRS International Journal of Geo-Information*, 2(3):817–836, 2013.

[143] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[144] D. J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of american Geographers*, 84(3):441–461, 1994.

[145] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 219–224. IEEE, 2005.

[146] Pharo. Pharo. `https://pharo.org/`, 2018. Accessed September 16, 2018.

[147] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '10, pages 120–123, Washington, DC, USA, 2010. IEEE Computer Society.

[148] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes. Do all birds tweet the same?: Characterizing twitter around the world. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1025–1030, New York, NY, USA, 2011. ACM.

[149] H. Purohit and A. P. Sheth. Twitris v3: From citizen sensing to analysis, coordination and action. In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[150] M. Quezada and B. Poblete. Understanding real-world events via multimedia summaries based on social indicators. In *Collaboration and Technology*, pages 18–25. Springer, 2013.

[151] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA, 2012. ACM.

[152] B. Robertson. "Fawcett" : A toolkit to begin an historical semantic web. *Digital*

Studies / Le champ numrique, 1(2), 2009.

[153] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, Nov 2008.

[154] A. Ronacher. Flask. `http://flask.pocoo.org/`, 2018. Accessed September 16, 2018.

[155] R. E. Roth. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2356–2365, Dec 2013.

[156] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. on Knowl. and Data Eng.*, 25(4):919–931, Apr. 2013.

[157] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, 2009.

[158] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[159] A. Saravanou, G. Valkanas, D. Gunopulos, and G. Andrienko. Twitter floods when it rains: A case study of the UK floods in early 2014. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15 Companion, pages 1233–1238, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[160] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *ICWSM*.

[161] A. Scherman, A. Arriagada, and S. Valenzuela. Student and environmental protests in chile: The role of social media. *Politics*, 35(2):151–171, 2015.

[162] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, Sep 1996.

[163] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[164] Sky News. Ukraine Protesters Now Want Leader's Head. `http://www.aljazeera.com/news/europe/2013/11/ukraine-drops-eu-plans-looks-russia-20131121145417227621.html`, 2013. Accessed November 14, 2017.

[165] F. M. Suchanek and N. Preda. Semantic culturomics. *Proceedings of VLDB Endowment*, 7(12):1215–1218, aug 2014.

[166] H. Sun and Z. Li. Effectiveness of cartogram for the representation of spatial data. *The Cartographic Journal*, 47(1):12–21, 2010.

[167] A. Swartz. Web.py. `http://webpy.org/`, 2018. Accessed September 16, 2018.

[168] Y. Tan, M. C. Vuran, and S. Goddard. Spatio-temporal event model for cyber-physical systems. In *Distributed Computing Systems Workshops, 2009. ICDCS Workshops '09. 29th IEEE International Conference on*, pages 44–50, June 2009.

[169] The Guardian. Malaysia airlines loses contact with plane carrying 239 people. `http://www.theguardian.com/world/2014/mar/08/malaysia-airlines-loses-contact-plane`, 2014. Accessed November 14, 2017.

[170] The Guardian. US sends investigators to join search for missing Malaysia Airlines plane. `https://www.theguardian.com/world/2014/mar/09/us-sends-investigators-fbi-malaysia-airlines-plane`, 2014. Accessed March 4, 2018.

[171] The Himalayan Times. All 8 bodies found at crashed US Marine chopper, Nepal army says. `http://www.foxnews.com/world/2015/05/16/all-8-bodies-found-at-crashed-us-marine-chopper-nepal-army-says.html`, 2015. Accessed 22 November 2016.

[172] The Himalayan Times. Japan assistance for Nepal earthquake recovery. `http://thehimalayantimes.com/business/japan-assistance-for-nepal-quake-recovery/`, 2015. Accessed 22 November 2016.

[173] The New York Times. Russia Sent Tanks to Separatists in Ukraine, U.S. Says. `http://www.nytimes.com/2014/06/14/world/europe/ukraine-claims-full-control-of-port-city-of-mariupol.html?_r=0`, 2014. Accessed November 14, 2017.

[174] The World Bank. `http://databank.worldbank.org/`, 2017. Accessed 7 September 2017.

[175] W. R. Tobler. A continuous transformation useful for districting. *Annals of the New York Academy of Sciences*, 219(1):215–220, 1973.

[176] D. Torres-Salinas, N. Robinson-García, E. Jiménez-Contreras, F. Herrera, and E. D. López-Cózar. On the use of biplot analysis for multivariate bibliometric and scientific indicators. *Journal of the American Society for Information Science and Technology*, 64(7):1468–1479, 2013.

[177] E. R. Tufte. Envisioning information. *Optometry & Vision Science*, 68(4):322–324, 1991.

[178] Twitter. Twitter API. https://developer.twitter.com/en/docs/api-reference-index.html, 2018. Accessed on May 12, 2018.

[179] C. Vaccari, A. Valeriani, P. Barberá, R. Bonneau, J. T. Jost, J. Nagler, and J. Tucker. Social media and political communication: a survey of twitter users during the 2013 italian general election. *Rivista italiana di scienza politica*, 43(3):381–410, 2013.

[180] M. van Kreveld and B. Speckmann. On rectangular cartograms. *Computational Geometry*, 37(3):175 – 187, 2007. Special Issue on the 20th European Workshop on Computational Geometry.

[181] Vatican Radio. Donors pledge billions of dollars to rebuild Nepal. http://en.radiovaticana.va/news/2015/06/25/donors_pledge_billions_of_dollars_to_rebuild_nepal/1153906, 2015. Accessed 22 November 2016.

[182] F. J. Villanueva, C. Aguirre, D. Villa, M. J. Santofimia, and J. C. Lpez. Smart city data stream visualization using glyphs. In *2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pages 399–403, July 2014.

[183] M. Walther and M. Kaisser. Geo-spatial Event Detection in the Twitter Stream. In *Advances in Information Retrieval*, pages 356–367. Springer Berlin Heidelberg, Berlin, Heidelberg, Jan. 2013.

[184] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *ECIR*, pages 356–367. Springer, 2013.

[185] X. Wang, W. Dou, W. Ribarsky, D. Skau, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 93–102, Washington, DC, USA, 2012. IEEE Computer Society.

[186] M. O. Ward. Multivariate data glyphs: Principles and practice. In *Handbook of data visualization*, pages 179–198. Springer, 2008.

[187] P. Warden. Geodict. https://github.com/petewarden/geodict, 2010. Accessed November 14, 2017.

[188] P. Warden. Data Science Toolkit. http://www.datasciencetoolkit.org/, 2011. Accessed November 14, 2017.

[189] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2541–2544, New York, NY, USA, 2011. ACM.

[190] Wikipedia. Comparative historical research. https://en.wikipedia.org/wiki/Comparative_historical_research, 2016. Accessed on November 3, 2017.

[191] Wikipedia. Israeli-Palestinian conflict. https://en.wikipedia.org/wiki/Israeli-Palestinian_conflict, 2016. Accessed November 23, 2017.

[192] Wikipedia. Ukranian crisis. https://en.wikipedia.org/wiki/Ukrainian_crisis, 2016. Accessed November 23, 2017.

[193] Wikipedia. April 2015 Nepal earthquake. https://en.wikipedia.org/wiki/April_2015_Nepal_earthquake, 2017. Accessed on November 10, 2017.

[194] Wikipedia. Chibok schoolgirls kidnapping. https://en.wikipedia.org/wiki/Chibok_schoolgirls_kidnapping, 2017. Accessed on November 13, 2017.

[195] Wikipedia. Quantitative history. https://en.wikipedia.org/wiki/Quantitative_history, 2017. Accessed November 3, 2017.

[196] Wikipedia. Regions of Chile. https://en.wikipedia.org/wiki/Regions_of_Chile, 2017. Accessed on November 10, 2017.

[197] Wikipedia. Disapearance of the Malaysia Airlines Flight 370, March 2014. https://en.wikipedia.org/wiki/Malaysia_Airlines_Flight_370, 2018. Accessed on January 18, 2018.

[198] Wikipedia. Yemeni Civil War. https://en.wikipedia.org/wiki/Yemeni_Civil_War_(2015%E2%80%93present), 2018. Accessed on January 18, 2018.

[199] L. Xie, H. Sundaram, and M. Campbell. Event mining in multimedia streams. *Proceedings of the IEEE*, 96(4):623–647, 2008.

[200] P. Xu, Y. Wu, E. Wei, T. Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, Dec 2013.

[201] Yahoo News. Ukraine PM resigns amid unrest, parliament revokes anti-protest laws. https://www.yahoo.com/news/ukraine-39-azarov-offers-resignation-government-press-083057414--sector.html, 2014. Accessed November 14, 2017.

[202] W. Yan. Ggebiplota windows application for graphical analysis of multienvironment trial data and other types of two-way data. *Agronomy Journal*, 93(5):1111–1118, 2001.

[203] N. Yusof, R. Zurita-Milla, M.-J. Kraak, and B. Retsios. Interactive discovery of sequential patterns in time series of wind data. *International Journal of Geographical Information Science*, 30(8):1486–1506, 2016.

[204] K. Zahra, F. O. Ostermann, and R. S. Purves. Geographic variability of twitter usage characteristics during disaster events. *Geo-spatial Information Science*, 20(3):231–240, 2017.

[205] C. Zhang, Y. Liu, and C. Wang. Time-space varying visual analysis of micro-blog

sentiment. In *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, VINCI '13, pages 64–71, New York, NY, USA, 2013. ACM.