WHERE IS HISTORY BEING WRITTEN? GEOREFERENCING CONTRIBUTIONS TO WIKIPEDIA

DAVID KALTSCHMIDT

Diplomarbeit

Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf
Institut für Informatik
Freie Universität Berlin

David Kaltschmidt: *Where is history being written? Georeferencing contributions to Wikipedia*
Diplomarbeit, © August 2011 – February 2012

SUPERVISORS:
Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf

LOCATION:
Written in the cafes of Berlin, Krakow, Paris, and Tallinn.

YEAR:
August 2011 – February 2012

PRODUCED WITH:
LaTeX ClassicThesis

In memory of my dad.

1951 – 2009

ABSTRACT

Wikipedia is more than an online encyclopedia. It is also a news channel as well as a self-updating history book. A global readership can follow political events as they unfold, written about by local people and later edited by other volunteers. This thesis describes a method to answer the question to what extent local volunteers write about events in their own country. First, the geographic origin of each individual article contribution is determined. In a second step, a given article is annotated with georeferences on a word level. The properties of these annotations then allow for a statistical geographic analysis. This analysis is performed on three example articles and then on a set of political articles from the English Wikipedia.

ZUSAMMENFASSUNG

Als Online-Enzyklopädie ist Wikipedia nicht nur Nachschlagewerk sondern auch ein sich stetig wandelndes Geschichtsbuch. Eine global verteilte Nutzerschaft liest und schreibt über lokale Ereignisse noch während sie passieren. Diese Arbeit beschreibt eine Methode zur Bestimmung des Anteils an Beiträgen, die vom betreffenden Land ausgehen. In einem ersten Schritt werden die geographischen Ursprünge aller Beiträge eines Artikels ermittelt. Mit den daraus erhaltenen Georeferenzen wird der Artikel Wort für Wort annotiert. Basierend auf diesen Annotationen kann dann der lokale Autoren-Anteil bestimmt werden. Es wird dann eine Autorenanalyse erst anhand von drei Beispiel-Artikeln und anschließend mit einer Gruppe von politischen Artikeln durchgeführt.

Sweat is what the gods put between us and virtue.

— Hesiod, Works and Days

ACKNOWLEDGMENTS

Above all, I would like to thank my mother for all her loving care over the course of my studies, for her unrelenting proofreading, and for all the shirts she will ever iron. I thank both my dad and my grandmother for instilling in my heart an insatiable appetite for the exploration of knowledge.

I am most grateful to my supervisor, Claudia, and the support and patience she has shown. I would also like to thank Molly Hannon and Peter Bourgon for all their input and rigorous views on copy-editing.

This thesis would not have been possible without the moral support of Beeke Urban, Lena Boers, and Söhnke Vosgerau, who kept me company during countless hours of studying. Many thanks also go to Andreas Heumann, Beata Biel, Dumisani Tekile, Liisi Sukles, Markus Düttmann, Sebastian Nehmsch, and Thomas Shuster, for taking my mind off this thesis once in a while, wherever they were located.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ABBREVIATIONS

API	application programming interface
CSS	Cascading Style Sheets
CSV	comma-separated values
HTML	HyperText Markup Language
IP	Internet Protocol
IQR	interquartile range
ISO	International Organization for Standardization

JS	JavaScript
JSON	JavaScript Object Notation
NPOV	neutral point of view
PHP	PHP: Hypertext Preprocessor
XML	Extensible Markup Language

Part I

THOUGHTS

1 INTRODUCTION

If you are open to contributions from others, you generally end up with richer, better, more diverse and expert content than if you try to do it alone.[1]

— Alan Rusbridger, editor of The Guardian

At the end of January 2011, when a wave of public protest spilled from Tunisia into Egypt, a small group of opposition parties and political activists called for a "Day of Rage" via Facebook, a social networking website. By January 25th their Facebook group had more than 80,000 supporters who drew attention to and helped organize the country-wide protests that followed. As people rallied the streets day after day, the Egyptian government first limited access to Twitter, a micro-blogging service, before cutting Egypt off the internet completely on January 28th.[2, 3]

In what came to be known as the Arab Spring, the use of online networks directly influenced the developing political situation. While Facebook played a part in organizing the protests, Twitter acted as an information channel during the demonstrations. As the events unravelled, they were reflected by articles created on Wikipedia, an online encyclopedia. Updated by the minute, the articles covering the protests formed a well of news reports.[4]

Wikipedia's free access, open editing policy, and high quality level—putting it "head to head"[5] with Encyclopædia Britannica—turned it into a hugely popular website[6]. The server software used for the website, MediaWiki[2], ensures that the effort to change an article is minimal. Given an internet connection and a web browser, anyone can add or edit an account of current events in a related article and publish it in a matter of seconds.

This form of news production turns the encyclopedia into a news channel that is constantly updated and corrected by an army of volunteers. The result is a self-governed news source that lends itself the aura of authority and credibility of a knowledge reference. At the same time, a technophile public that uses

1 The Economist. *The people formerly known as the audience*. URL: http://www.economist.com/node/18904124 (visited on 08/10/2011)

2 http://www.mediawiki.org (visited on 10/31/2011)

the internet as an efficient means of news acquisition can check facts on Wikipedia, and act upon the consumed information.[7, p. 424–427] Therefore the collective authorship of such a news medium could have a direct influence on the political decision-making process. As ordinary people become producers of journalism, the need arises to analyze these contributions. This thesis investigates how Wikipedia is being used during and after political events with a special focus on the geographic origins of contributions to Wikipedia articles treating those events.

Political events are often limited to a country or region. This is reflected by the Wikipedia articles covering the Arab Spring that will serve as examples throughout this thesis. There is an overarching parent article *Arab Spring*[3] as well as single articles covering the revolution in each of the affected countries, e.g. Egypt[4], Libya[5], and Bahrain[6]. The events in Libya also exemplify how divided the political actors can be. While nearly all revolutionaries welcomed the airstrikes, one faction was concerned about foreign meddling and another one just opposed the deployment of ground troops.[8]

The collective authorship could be equally divided. Despite Wikipedia's core policy to oblige everyone to write from a neutral point of view (NPOV)[7], people regularly express opinions. The collision of opinions in a collectively written article can result in a prolonged series of edits and subsequent reversals by conflicting parties. The resulting edit pattern is known as an *edit war*.[9] These clashes create a potential for further investigation into the geopolitics of article contribution. Where do the first reports of an event originate? As later iterations of revisions turn these reports into historical accounts, are these editors from the same country? And more generally, to what extent is a collection of these articles written by volunteers located at the respective location of the event?

In this thesis I will propose a method to help answer these questions. First, I will present an improved technique for determining the geographic origin of contributions to an article. This georeferencing is a key step in the analysis of article revisions. Then, from the geographic distribution of contributors,

3 http://en.wikipedia.org/wiki/Arab_Spring (visited on 01/29/2012)
4 http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011 (visited on 10/31/2011)
5 http://en.wikipedia.org/wiki/2011_Libyan_uprising (visited on 10/31/2011)
6 http://en.wikipedia.org/wiki/2011_Bahraini_uprising (visited on 01/29/2012)
7 http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view (visited on 12/08/2011)

local and distant participation will be determined for individual articles. I will also present two variations of a proximity metric that will allow to measure whether local contributions are more likely to prevail over time. These metrics will also be applied in a quantitative analysis of a group or political articles.

1.1 STRUCTURE

The chapter FOUNDATION provides background information about Wikipedia, article editing (Contributions), and the application of geographic data (Georeferences). The first part ends with the HYPOTHESES where I propose the research questions that this thesis hopes to answer.

In APPARATUS I will describe the data sources and tools that I used in the development of a web application to perform the revision analysis. The chapter details the improved user-locating algorithm (Locating users) as well as the variations of the proximity metrics (Signature distance). In EXPERIMENTS, I will present the results of an analysis for a few example articles. There, I will describe how the application's various types of visualization can be used to interpret the results.

Finally, in RESULTS I will define the data set for the quantitive analysis and present the results of the application run. Their feasibility as well as the method's limitations will be discussed in CONCLUSION.

2 FOUNDATION

Wikipedia is a phenomenon that has attracted researchers across all fields, notably computer science and sociology, who have written over 1,000 reports on the subject to date. Nielsen [10] compiled an overview of Wikipedia research[1] and divides these publications into four categories:

CONTENT PRODUCTION Covering all aspects of voluntary production such as motivation, collaboration, coverage and bias, quality and vandalism, actuality, and geography.

INFORMATION USE Treating how the resulting corpus is being used, e.g. Wikipedia citation in research, use in court, trend-spotting, natural language processing and automatic translation tools, thesaurus construction, or categorization.

IMPROVEMENT These are studies concerned with the improvement of both the software used by Wikipedia and the content, e.g. automatic linking, improved editors, as well as quality and trust indicators.

COMMUNICATION Studies in this category look at Wikipedia as an online collaboration tool for education and research.

This thesis falls into the first category, content production, as it examines the geography of article contributions that will become part of the Wikipedia's corpus. After a short overview of Wikipedia from a user's perspective, I will introduce its model of collective authorship and present prior research concerning location and geography.

2.1 WIKIPEDIA

Wikipedia is an online encyclopedia with editions in over 260 languages. Counting 3.6 million articles, the English version is by far the biggest. However, other language editions differ sharply in size and usage.[11] If articles covering the same topic

1 Another resource is the Wikimedia Foundation's own directory of Wikipedia research projects at http://meta.wikimedia.org/wiki/Research:Projects (visited on 10/12/2011).

exist in other language editions, they can be connected by inter-language links.[2] The thesis will focus on the English edition.

2.1.1 *History*

Wikipedia was officially started on 15 January 2001 by Jimmy Wales and Larry Sanger. Wales previously founded Nupedia, a free and peer-reviewed online encyclopedia written only by experts. However, the speed of content production was extremely low. Wikipedia was founded as a feeder project to collectively write on articles before these entered Nupedia's review process. Wikipedia then quickly created other language editions and overshadowed its predecessor[3].[12]

After being mentioned on Slashdot, a technology news website, in March 2001, Wikipedia quickly attracted new users. This tech-savvy group of people created new articles at a staggering rate of 1,500 articles per month in the first year. These articles then quickly started showing up in Google's search results, attracting even more new users. The non-English editions grew slower but as a group accounted for 75% of all articles in 2007. By 2011 the combined article count passed 20 million.[12]

2.1.2 *Wikimedia Foundation*

Wikipedia is operated by the Wikimedia Foundation, a non-profit organization founded in Florida on June 20, 2003. It is completely financed by public contributions, such as donations and grants. Individual grants can be quite substantial—among the most generous donors are Google and the Stanton Foundation handing out respectively $2 million and $3.6 million in single donations.[13]

In the Wikimedia Foundation's 2010–11 fiscal year, $8.9 million was spent on website operations, including server hosting and software maintenance. The rest of the $20.0 million of total expenditures went into complementary activities such as fund raising, administration, and the support of local chapters.[14]

The local chapters are self-dependent organizations set up in countries around the globe to locally promote the foundation's cause and collect donations. The first local chapter to be created was Wikimedia Deutschland, founded in Berlin in 2004.[13]

2 As these links are in fact links between different wikis, they are also called *interwiki* links.

3 Only 24 articles were completed in Nupedia's review process. The project was officially ended in 2003.

The individual language editions of Wikipedia are not hosted by the local chapters, however. All of Wikipedia's content is centrally stored on servers in Tampa, Florida and in Amsterdam, Netherlands.[13]

2.1.3 *Anatomy of an article*

All Wikipedia articles share a similar layout: a large content area topped by the article title. Article titles can change over time, e.g. *2011 Bahraini uprising* was renamed to *2011-2012 Bahraini uprising*[4]. For these cases, Wikipedia has a redirecting mechanism that forwards the visitor to the final article and displays a small note below the title (see figure 1).

Figure 1: The article *2011–2012 Bahraini uprising* viewed in a web browser on 01/23/2012.

Occasionally the content section can be topped by one or more warning boxes to inform the visitor that the article is violating an editing policy, e.g. the information of the article may be outdated because it is subject to current events. When an article spans several sections a table of contents is added below the first introductory paragraphs. In addition to prose, some articles feature info boxes on the right hand side. Theses boxes show information in a structured way and can be found on articles of similar topics, giving the visitor a quick glance on key information without having to read the text.

This information may include dates and geographic coordinates. E.g. the article *2011–2012 Bahraini uprising* has the time interval "14 February 2011 – ongoing" and is tagged with the coordinates 26°01′39″N 50°33′00″E, pointing to the centre of Bahrain. Even when no coordinates are present in the article, it still may

4 http://en.wikipedia.org/wiki/2011-2012_Bahraini_uprising (visited on 01/23/2012)

be associated to a location. In that case the info box just presents the place, instead of providing the geographic coordinates, e.g. *Maspero demonstrations*[5].

2.1.4 *Categories*

At the bottom of each article is an optional list of categories that the article belongs to, e.g. the article *2011-2012 Bahraini uprising* belongs, among others, to "Arab Spring by country"[6] and "2011 protests"[7].

Categories can not only consist of pages but also of subcategories, e.g. "Arab Spring by country" has the subcategory "2011 Libyan civil war"[8] which in turn has 5 sub-categories and 55 pages. When looked at as a graph, the system of categories does not form a tree, however, for there is no restriction on the inclusion of categories—even cycle-creating inclusions are possible. The set of articles in a category can be as arbitrary as its topology. The category "Arab Spring by country" does not only contain articles covering the Arab Spring by country, but also articles about killed individuals, e.g. "Zakariya Rashid Hassan al-Ashiri"[9].

2.2 MEDIAWIKI AND EDITING

Anyone with a browser and internet access can edit Wikipedia's articles[10]. In collaboration, people all over the world contribute and improve the content. This is made possible by MediaWiki, the server software that makes Wikipedia a wiki. The software allows website visitors to add and modify the page content in the browser using *wikitext*, simplified markup language.[11] Its syntax can be used to structure a text into sections, embed images, and links to other pages, similar to HyperText Markup Language

5 http://en.wikipedia.org/wiki/Maspero_demonstrations (visited on 01/25/2012)
6 http://en.wikipedia.org/wiki/Category:Arab_Spring_by_country (visited on 01/23/2012)
7 http://en.wikipedia.org/wiki/Category:2011_protests (visited on 01/23/2012)
8 http://en.wikipedia.org/wiki/Category:2011_Libyan_civil_war (visited on 01/23/2012)
9 http://en.wikipedia.org/wiki/Zakariya_Rashid_Hassan_al-Ashiri (visited on 01/23/2012)
10 Some articles can be locked because of sustained vandalism or content disputes.[15]
11 For the syntax see http://en.wikipedia.org/wiki/Help:Wiki_markup (visited on 12/12/2011).

(HTML). The syntax is explicitly kept simple to keep the entry barrier to editing low, e.g. adding an article to a category is as easy as putting the category name at the end of the wikitext.

2.2.1 *Templates*

Wikitext has a special syntax for templates[12]. These are reusable containers for text snippets and repetitive material like the info boxes described in Anatomy of an article. When a template is used in a page, the server software replaces the template placeholder—the template name surrounded by curly brackets—with the template's content. The content can be parameterized with key-value pairs so that, for example, an info box about countries can be used by several countries' articles.

By using templates, the information is likely to be more structured than simple, free-form text. Each invocation of a template also renders its content in the same way, allowing for and encouraging more consistency[13]. More importantly, the usage of a template in an article lets that article become a member of the group of articles that embed this template. This is an alternative mechanism to group articles that is likely to yield more homogenous results than the category system, see Categories.

2.2.2 *Revision history*

Each submission of an edit in the browser creates a new revision of the article and is stored in the revision history, see figure 2. Naturally, each article available today started from an empty page and is the result of a succession of edits.

Each entry in the revision history consists of the new wikitext, the date of submission, the user name, and an optional comment explaining the change. In addition, authors have the possibility to mark an edit as *minor*. In doing so, they suggest that the submitted change is only superficial, e.g. correcting typography or moving passages of text, and does not change the meaning of the article.[16]

Each revision can not only be examined by other users but also reverted. Especially in the case of vandalism this mechanism can be used to restore the previous state of the article. Reverts may also appear when different views on the same topic collide. To

12 http://en.wikipedia.org/wiki/Help:Template (visited on 01/23/2012)

13 E.g. when an editor notices that an info box supports the parameter "location" but does not have one yet, the user may feel encouraged to complete it.

Figure 2: Revision history of the article *2011-2012 Bahraini uprising* on 01/23/2012.

minimize the potential for *edit wars*[9] Wikipedia urges its users to discuss controversial topics on the article's talk page.

2.2.3 *Authors*

Contributions to an article can be done anonymously or as a registered user. A registered user gains privileges like the ability to create articles or the use of the social network features in Wikipedia. With the initial registration a *user page* is created where the user is allowed to publish a profile and interact with other registered users.[17] The majority of edits come from registered users; anonymous edits account for a quarter of all edits.[18]

A third group of editors are automatic programs known as *bots*. They perform routine tasks ranging from spell-checking, to curse word detection, to automatic reverts on vandalism. Currently the English Wikipedia alone has nearly approved 1,500 bot tasks running, either automatically or manually triggered by a real user.[19]

2.2.4 *User pages*

When a Wikipedia user decides to register, a *user page* is created for him on Wikipedia's website. This is a special page that can be edited like any other article. The user can publish personal information in prose, or reuse a template (see Templates). The templates available to decorate one's user page include the following:

- Spoken languages, e.g. "This user is a native speaker of English."[14]
- Location, e.g. "This user comes from India."[15]
- Expression of personal views, e.g. "This user opposes Imperialism."[16]

Of course, a user can publish just about anything. As a result, the information on a user page may not be accurate.[17]

Like any other article, each user page has a discussion page that can be used to communicate with that user by leaving a message. Viégas et al. [20] looked at how contributors coordinate their actions and found that these pages "hold much of the community interaction".

2.3 CONTRIBUTIONS

Wikipedia's articles are continuously edited by its users. The nature of an edit can range from simple spelling or grammar correction, to improving the content of a sentence, to writing or removing whole articles. This collective authorship makes it difficult to determine an individual author's contributions; in other words, it is not easy to tell who wrote what.

Research in this area tends to be motivated by the desire to identify individual authors with a good reputation in order to assign a trust score to them. This is based on the assumption that trusted authors consistently produce high quality contributions that outlive contributions of lower quality. Kramer, Gregorowicz, and Iyer [21] devised a method to assign trust scores to the authors of an article by examining the wealth of information contained in the article's revision history. They looked at an article as being a set of phrases. The author who first wrote a sentence gets the credit for that phrase and will gain trust if it survives future edits.[18]

14 http://en.wikipedia.org/wiki/Wikipedia:Babel (visited on 01/23/2012)

15 http://en.wikipedia.org/wiki/Template:User_India (visited on 01/23/2012)

16 http://en.wikipedia.org/wiki/User:Serouj/UserBox/Against_Imperialism (visited on 01/23/2012)

17 See user Lihaas, who seems to hail both from India and from Pakistan: http://en.wikipedia.org/wiki/User:Lihaas (visited on 01/23/2012)

18 Kramer, Gregorowicz, and Iyer [21] define a sentence as an n-gram—a sequence of n words—and use a sliding window model to follow it across revisions to prevent simple rearrangements of text from counting as a new sentence.

A similar approach of calculating the longevity of text chunks was followed by Adler and De Alfaro [22]. They adapted standard text-diff algorithms to the peculiarities of the wiki revision system, e.g. keeping track of text chunks that were removed at one point and then reinserted in later revisions. Based on these algorithms a reputation system was implemented by Adler et al. [23] which offers an application programming interface (API)[19] that can annotate a Wikipedia article. The annotated text is the result of splitting the original text into chunks and attributing them with their respective authors, the number of the revision where the chunk was added, and a trust value for the author.[20]

2.4 GEOREFERENCES

In order to analyze the localness of contributions, it is necessary to geotag them, i.e. applying geospatial metadata like coordinates to each contribution, derived from the author's location. In his doctoral thesis Hardy [24] used the Wikipedia corpora to study the spatial behavior of article production. The dataset was limited to anonymous users and articles that were geotagged.

For each anonymous contribution an Internet Protocol (IP) address, belonging to the point of Internet access, is stored in the revision that is created. Various methods to determine the geographic location from a given IP address have been studied by Muir and Oorschot [25]. Various visualizations[21] [22] of edit distributions use geolocation databases like MaxMind[23] and Quova[24].

For registered users, the IP address is not stored with the revision. Therefore IP geolocation services cannot be used. Lieberman and Lin [26] found a novel approach by assuming users prefer to edit geographic articles in their proximity. The approximated user location was derived from the center of the convex hull around those articles.

The user pages offer another source for finding the author's location. Entity names like a city or a country can be extracted from both the prose or the info boxes that the users put on their page.

19 http://www.wikitrust.net/vandalism-api (visited on 10/31/2011)
20 Adler et al. also released a Firefox add-on that highlights untrustworthy passages when viewing Wikipedia articles: https://addons.mozilla.org/en-US/firefox/addon/wikitrust/ (visited on 11/15/2011).
21 http://infodisiac.com/blog/2011/05/wikipedia-edits-visualized/ (visited on 10/31/2011)
22 http://sonetlab.fbk.eu/wikitrip/ (visited on 10/31/2011)
23 http://www.maxmind.com (visited on 10/31/2011)
24 http://www.quova.com (visited on 10/31/2011)

Building on this foundation, I will formulate a set of HYPOTHESES in the next section.

3

HYPOTHESES

To gain insight on how Wikipedia is being used during and after political events[1], and ultimately whether articles covering the events are written by people that are most affected, I will propose a set of hypotheses aimed at different aspects of article production.

For this thesis I will use the event definition proposed by Lewis [27, p. 243]:

> "An event is a localized matter of contingent fact. [...] An event occurs in a particular spatiotemporal region."

It follows that events must have clear spatial and temporal boundaries. The spatial boundaries give it a location, distinguishing the where. In addition to the innate temporal boundaries, namely the start and the end date[2], I will further denote the *beginning* of an event as the first seven days of an event.[3] Contributions that occur in the beginning will be referred to as *early*. The rest of contributions appearing in the unbounded time interval after the beginning will be referred to as *late*.

Let me further denote authors in the event's proximity[4] as *local*, while the opposite be referred to as *distant*.

The observation of both the localness and the division into time intervals will allow for a detailed look into Article creation, the level of Participation, as well as Text survival.

3.1 ARTICLE CREATION

The use of Wikipedia during a political event starts with the creation of the article describing the event. How quickly users create such articles is less clear. Wikipedia's popularity and global

1 I will not try to argue what makes an event political but rather identify a set of events by picking suitable categories of articles.

2 For ongoing events the end date will be the date of the analysis.

3 The interval was picked arbitrarily but acknowledges the fact that event dates in Wikipedia articles rarely carry a time attribute, therefor a shorter interval, say 24 hours, is less feasible.

4 Hardy [24, p. 57] defined proximity not in absolute terms, rather he considered the likeliness of authors being located less far than the average distance between an article and all its contributors. In the next chapter APPARATUS I will describe the algorithm used to distinguish local from distant authors.

reach provide an incentive to use it as a news channel and to publish without delay. This leads to the following hypothesis:

HYPOTHESIS 1 (H1) Articles are created shortly after the start of the event.

Moreover, considering the rising year-on-year Wikipedia usage in numbers of pages viewed[6], the incentive becomes greater. Consequently, for more recent events the delay between the start date and the creation of the article should become shorter. This would also suggest an increased use of Wikipedia as a news channel, thus:

HYPOTHESIS 2 (H2) The more recent an article, the shorter is the delay between the event start and article creation.

A user has the chance to create a new article in any of the 260-odd language editions of Wikipedia. Although the English version is by far the biggest and most visited, it would be interesting to see whether it is also the first choice to create an article covering a new event. Shortly after the first article has been created, articles covering the same topic are produced across various editions of Wikipedia. These articles are then being linked, mostly manually, via interlanguage links.[5] When studying knowledge diversity across language editions, Hecht and Gergle [29] found that the English edition is not the superset of concepts of all editions as was previously believed. This means authors retain knowledge they consider important only to their compatriots. For a citizen of a country where English is not an official language, creating an article becomes a political decision: should the author make the information available to fellow citizens or to a world-wide readership? The following hypothesis will be tested only for articles where the user faces that choice, i.e. for articles about events in countries where English is not an official language, thus:

HYPOTHESIS 3 (H3) Articles are being created first in the English Wikipedia.

Regarding the localness of contributions, Hardy [24, p. 57] has established that Wikipedians write about places in their proximity more often than distant ones. His sample included only

5 Adar, Skinner, and Weld [28] found that between two languages the interlinking is not symmetrical, i.e. the number of out-links does not match the number of in-links. Links are either missing on one side or the respective topics are not congruous and the user intentionally left out one direction.

articles that have a geotag. Naturally, articles about geographic places like towns and sites[6] will dominate this sample. Since my thesis is concerned with political events, I find this point to be worth revisiting[7]. People in an event's proximity are probably more likely to be affected by it. The following hypothesis aims to find out whether they are the ones creating the article, thus:

HYPOTHESIS 4 (H4) Articles about political events are created by local authors.

Hypotheses 1–4 will only be tested against articles that were created as a reaction to an event that has already started. This excludes scheduled events like elections, e.g. *Russian legislative election, 2011*[8] which was created 335 days before the election date, almost a year in advance.

3.2 PARTICIPATION

A Wikipedia article usually has more than one author. Once it has been created, users from around the globe can edit an article collectively. Viégas, Wattenberg, and Dave [31] tried to find patterns in the revision history that would reveal certain aspects of collaboration or the lack thereof, e.g. discussions and vandalism.[9] In respect to authorship, the researchers found the proportions of anonymous contributions differed strongly from page to page while showing no preference to any topic. This inconclusive result and the age of the sample[10] merits further investigation.

In 2007, Kittur et al. [32] found that a core of registered users is still doing the bulk of all edits. However, anonymous users contribute considerable amounts of text. For accounts of political events, due to their dynamic nature, I expect a strong participation by unregistered users while events are still unfolding:

HYPOTHESIS 5 (H5) In the beginning most contributions are anonymous.

6 According to Kittur, Chi, and Suh [30] articles about "geography and places" are third biggest group.

7 In addition, Hardy [24, p. 61] considered only anonymous users. Since creating an article is only allowed for registered users, his method has to be extended.

8 http://en.wikipedia.org/wiki/Russian_legislative_election,_2011 (visited on 01/07/2012)

9 Using their history flow visualization Viégas, Wattenberg, and Dave [31] first identified patterns in single articles and later tried to statistically confirm their prevalence by analyzing the complete English corpus.

10 Viégas, Wattenberg, and Dave [31] used a dataset from May 2003.

Furthermore, Viégas, Wattenberg, and Dave [31] found that early contributions have a high survival rate. Recognizing this *first-mover advantage*, I suspect that accounts of political events show a strong localness in the beginning:

HYPOTHESIS 6 (H6) In the beginning most contributions are local.

After the beginning, I expect these parameters to change. Looking at the whole lifespan of an article I would also expect registered users to outnumber anonymous ones as suggested by Kittur et al. [32], thus:

HYPOTHESIS 7 (H7) Later, the share of anonymous contributions decreases over time.

Similarly, I expect distant authors to participate more as the global readership joins the authors in contributing to the article. These additional contributions by distant authors lessening the proportion of local contributions:

HYPOTHESIS 8 (H8) Later, the share of local contributions decreases over time.

3.3 TEXT SURVIVAL

In Participation, the edits are all treated as a single unit of contribution. This gives authors with a higher edit count more credit. However, when multiple authors write the same article, they do not only add text but also modify or even delete parts. In some cases, contributions are reverted immediately, i.e. the contributed text is no longer part of the article. Recognizing this dynamic, the last two hypotheses are concerned with the relative performance of local contributions:

HYPOTHESIS 9 (H9) Local contributions are more likely to survive.

HYPOTHESIS 10 (H10) Text from local contributions is more likely to survive.

If support for H10 can be found, text written by locals will be overrepresented in an article. This concludes the statement of the hypotheses. The next chapter, APPARATUS, describes an application designed to test these.

Part II

METHODS

4

APPARATUS

This chapter describes the application I designed to test the hypotheses. The first section lists the Data sources and shows what type of content they provide for the article analysis. In Application design, I will give an overview of the application's architecture as well as the technologies being employed.

The section Algorithms describes the article requirements to qualify for an analysis. More importantly, the same section features the algorithms for the proximity metric, locating users, and information extraction, e.g. date and location parsing.

In Visualization I give a quick introduction into the charts being used to present the data, and that will be featured more prominently in the following chapter EXPERIMENTS.

The application is can be accessed over the internet[1]. Its source code is hosted on GitHub[2].

4.1 DATA SOURCES

For an automated analysis, simply browsing Wikipedia's website is not really feasible. The bulk of Wikipedia's content, e.g. articles, revisions, and discussions, is stored on its database servers. Unfortunately, these databases are not directly accessible over the Internet. The Wikimedia Foundation, however, makes a lot of the data available in the form of database dumps or through an API.

4.1.1 *Wikipedia website*

For a complete article analysis, navigating the website can be tedious, as one would have to click through a complete revision history and parse the page's source which is formatted in HTML. However, individual pages contain data that is static and can be used throughout the analysis process. This makes it worth writing a specific parser for a technique known as *screen scraping* to extract the information. On a high level, it involves the following steps:

1 http://davkal.github.com/WP-contributions/ (visited on 02/04/2012)
2 https://github.com/davkal/WP-contributions (visited on 02/04/2012)

1. Looking at the HTML source of the page and identifying how the HTML tags and attributes are used to structure the information.

2. Writing a parser that addresses the identifying tags and thereby tokenizes the data.

3. Converting the found tokens into an output format, e.g. JavaScript Object Notation (JSON).

For a simple HTML table, a parser can be written in a few lines of code. Using this technique, the following static information was gathered:

BOTS A list of bots was built based on the Wikipedia page *List of bots by number of edits*[3]. This list is used to distinguish bots from real authors, as contributions done by bots are excluded from the analysis. There are unregistered bots, however, which do not appear on the list. For a lack of automated distinction, these are counted as normal authors.[4]

COUNTRIES A list of countries was extracted from the article *ISO_3166-1*[5]. It provides a list of standardized country names that is also respected by Wikipedia's authors when referring to a country by name. In a second pass, the Wikipedia article of each country was retrieved to extract the country's geographic coordinates[6]. For a discussion on countries and coordinates, see Geographic resolution.

OFFICIAL LANGUAGES For each entry in the list of countries, the country's official languages were added. From the *List of ISO 639-1 codes*[7], all languages including their two-letter codes were extracted. In a second step, the *List of official languages by state*[8] was scraped and merged with the first

3 http://en.wikipedia.org/wiki/Wikipedia:List_of_bots_by_number_of_edits (visited on 01/24/2012)

4 A simple heuristic employed by other software to analyze MediaWiki content is treating all contributors whose username contains or whose comments start with "bot" as a bot, e.g. pymwdat, see http://code.google.com/p/pymwdat/source/browse/trunk/toolkit.py?spec=svn13&r=13 (visited on 01/24/2012). This has a potential for false positives and is not used.

5 http://en.wikipedia.org/wiki/ISO_3166-1 (visited on 01/02/2012)

6 For some countries, coordinates were not present on the page, e.g. http://en.wikipedia.org/wiki/Australia (visited on 01/24/2012). In that case, they were manually added by using that country capital's coordinates.

7 http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes (visited on 01/29/2012)

8 http://en.wikipedia.org/wiki/List_of_official_languages_by_state (visited on 01/29/2012)

list by matching the language name.[9] The result is a mapping from a country's name to a list of two-letter codes of that country's official languages.

Making both of these sets static was a design decision recognizing the trade-off between having them in memory and querying for each article.

4.1.2 *Database dumps*

Monthly database snapshots of all wikis run by the Wikimedia Foundation, including Wikipedia, are publicly available[10] as database dump files in the Extensible Markup Language (XML) file format. For each of the wikis a variety of dumps is available that include all articles and, optionally, their revision history, all categories, interlanguage links, etc. Despite this openness, some database tables are not publicly available. The dump files of the database tables *users* and the *watchlist* are kept private.

The dump files can be quite large, e.g. a compressed dump of all articles of the English Wikipedia in their current revision has a size 7.3 GB.[11] This huge size makes processing them rather slow.[12] When analyzing only a single article or a category of articles, the MediaWiki API can deliver the same information contained in the dumps in a much more targeted manner.

4.1.3 *MediaWiki* API

Wikipedia runs on the open source software MediaWiki, written in PHP: Hypertext Preprocessor (PHP). It offers a well documented API[13] which can be used by other programs to remotely use the wiki's features such as changing content and restoring revisions.[14] For analysis of articles, the API offers queries directed at a variety of article properties, e.g. revisions, categories and,

9 Not all countries have official languages—including the United States of America. In that case, the *de facto* official languages were used, e.g. English and Spanish for the United States of America.

10 http://dumps.wikimedia.org (visited on 12/11/2011)

11 The uncompressed size is 31.0 GB, see http://en.wikipedia.org/wiki/Wikipedia:Database_download (visited on 12/11/2011)

12 The project WikiHadoop addresses this problem by offering a stream task format to be used in Hadoop (MapReduce) infrastructure, see https://github.com/whym/wikihadoop (visited on 12/11/2011).

13 http://www.mediawiki.org/wiki/API:Main_page (visited on 01/24/2012)

14 The full capability of the API can be seen at tried at the *Sandbox* at https://en.wikipedia.org/wiki/Special:ApiSandbox (visited on 01/24/2012), a recent addition to the MediaWiki software.

links. Among the output formats for the responses are JSON and XML. Similar to MediaWiki's Special:Export page[15], the API also offers an article export that includes all revisions.

Listing 1: Example JSON response to a query to list all bots that edited the article *2011-2012 Bahraini uprising*

```
{
    "query": {
        "redirects": [{
                "from": "2011 Bahraini uprising",
                "to": "2011-2012 Bahraini uprising"
        }],
        "pages": {
            "30876395": {
                "pageid": 30876395,
                "ns": 0,
                "title": "2011-2012 Bahraini uprising"
            }
        },
        "allusers": [{
                "userid": "13146235", "name": "28bot"
            }, {
                "userid": "5415725", "name": "718 Bot"
            }, ..., {
                "userid": "13770078", "name": "AWBCPBot"
        }]
    },
    "query-continue": {
        "allusers": {
            "aufrom": "AWeenieBot"
        }
    }
}
```

Some of the queries have a limit on how many results they return on a single request. When there are more results, the response contains a *query-continue* attribute that can be sent with following query so that the next result set can be returned. The following API calls will be important for this thesis:

QUERY INFO This basic query returns essential information like the article ID, the last revision ID, but also the full wikitext of the last revision.

QUERY REVISIONS Lists all revisions for an article and for each includes a timestamp, and the user, as well as the comment for the text change.

15 The page https://en.wikipedia.org/wiki/Special:Export (visited on 12/11/2011) allows for exporting of articles from the English Wikipedia.

QUERY CATEGORYMEMBERS For a given category, this query lists the articles and subcategories that belong to it. This query can be used to construct groups of articles for analysis in this thesis.

QUERY EMBEDDEDIN For a given template name, this query lists all pages that embed it. This query can also be used to build a group of articles.

OPEN SEARCH A method to suggest articles, categories, and templates that contain a term. It can be attached to an input field where a user is supposed to enter the name of an article.

PARSE This special query returns the HTML version of the article's wikitext. The content that is returned is exactly the HTML source that is sent to a browser when a visitor looks at this article or user page. This query will be used in cases where it is easier to parse the HTML markup than the wikitext, e.g. the User pages.

4.1.4 *Toolserver*

The Germany based Wikimedia Deutschland e.V. runs Toolserver[16], a platform for software tools that can access a continuously updated copy of Wikipedia's databases. Among these replicated databases is the English Wikipedia and other major language editions. However, the deployment of self-made software scripts is restricted and requires an account on Wikimedia's Toolserver.[17]

Some scripts that are already deployed can be accessed freely, allowing them to be reused. One of these was developed by SoNet[18], a social networking research group based in Italy, for a project called WikiTrip (see Analysis projects). It offers an API[19] to get simple article statistics like article ID, text length, as well as complex data structures like a list of unique editors including

16 http://toolserver.org (visited on 12/11/2011)

17 I applied for a Toolserver account outlining my necessary database queries, usage profile, as well as my affiliation with the Freie Universität Berlin. The application was submitted on 2011-12-21 and has not been processed yet (2012-01-23).

18 http://sonetlab.fbk.eu/ (visited on 12/12/2011)

19 The API is documented here: https://github.com/volpino/toolserver-scripts/tree/master/\ac{PHP} (visited on 12/12/2011)

their gender, if they are registered users and chose to reveal their gender in their Wikipedia account.[20]

In effect, calling the SoNet API replaces several calls to the original MediaWiki API and thereby speeds up the information retrieval, especially when the number of revisions or authors is high. The returned data object has the following structure:

Listing 2: SoNet API response to a query for the article *2011-2012 Bahraini uprising*

```
{
    "first_edit": {"timestamp":1297734917,"user":"Master&Expert
        "},
    "count":1778,
    "minor_count":401,
    "count_history":{"today":3,"week":5,"month":90,"year"
        :1778},
    "last_edit":1327370324,
    "totaldays":0,
    "average_days_per_edit":"0.00",
    "edits_per_month":0,
    "edits_per_year":0,
    "edits_per_editor":"4.17",
    "editor_count":426,
    "anon_count":337,
    "editors": {"Bahraini Activist":
        {"all":106,"minor":21,"first":"17 May 2011, 09:45:25",
        "last":"22 January 2012, 10:52:50","atbe" 203811,
        "minorpct":"19.81", "size":"140.54","urlencoded":"
            Bahraini_Activist"},
        ...
    },
    "anons":{"2011-02-15T08:11:52Z":
        ["78.2.29.139","Rovinj Croatia",45.08,13.64],
        ...
    }
}
```

This high density of preprocessed information shows the power of the Toolserver and its direct access to the database. The property *editors* lists all unique authors of an article and their edit count (property *all*). The second exhaustive collection is under the property *anons*. There, all anonymous contributors are listed with their IP addresses, as well as their geographic regions and coordinates. The geographic lookup uses[21] the GeoCityLite database from Maxmind (see IP Look-up for a discussion).

20 Try http://toolserver.org/~sonet/api_gender.\ac{PHP}?article=Egypt&lang=en (visited on 12/11/2011) to get a list of all registered users who edited the article *Egypt* of the English Wikipedia.

21 https://github.com/volpino/toolserver-scripts/blob/master/\ac{PHP}/api.\ac{PHP} (visited on 01/24/2012)

4.1.5 *Third-party sources/web services*

Like the Toolserver scripts in the previous section, other research projects exist that can be reused as data sources. Depending on the project's goal, a variety of preprocessed data is available:

ARTICLE TRAFFIC Wikipedia user Henrik[22] provides a web service that processes Wikipedia's log files[23] to calculate the number page views per article for a given time. These statistics can be viewed through a browser[24] or queried through an API [25].[26]

CATSCAN This web service, offered by Toolserver administrator Duesentrieb[27], finds articles that belong to a given category and its sub-categories (see Categories on why this is non-trivial). It also offers to limit the search to an intersection of categories, e.g. German politicians who are also physicists[28]. The results are presented in the browser or can be downloaded as a file in the comma-separated values (CSV) format.[29]

POOR MAN'S CHECKUSER The project *Poor Man's Check User*[30] mapped registered users to IP addresses based on a bug in the session management of the MediaWiki software.[31] For

22 http://en.wikipedia.org/wiki/User:Henrik (visited on 12/12/2011)

23 These are available at http://dumps.wikimedia.org/other/pagecounts-raw/ (visited on 12/12/2011)

24 E.g. http://stats.grok.se/en/201105/2011_Egyptian_Revolution (visited on 12/12/2011)

25 E.g. http://stats.grok.se/json/en/201105/2011_Egyptian_Revolution (visited on 12/12/2011)

26 I built support for this data source into the application. It is disabled, however, as page views are out of scope regarding the content analysis.

27 http://meta.wikimedia.org/wiki/User:Duesentrieb (visited on 12/13/2011)

28 https://toolserver.org/~daniel/WikiSense/CategoryIntersect.\ac{PHP}?wikilang=de&wikifam=.wikipedia.org&basecat=Politiker+(Deutschland)&basedeep=3&mode=cs&tagcat=Physiker&tagdeep=2 (visited on 12/13/2011)

29 This data source is not available in my application.

30 Project website: http://wikiwatcher.virgil.gr/pmcu (visited on 01/02/2012). The project's name is a reference to the *checkuser* permission that a community-elected group of registered users possesses. It allows a de-masking of the IP addresses for each of a registered user's edit.

31 When a user exceeded a certain time while editing an article without submitting the current changes, the user's session expired on the server. When the edit was submitted after the expiration the user appeared as an anonymous author, being only known by his IP address. When the user then logged in again, the same change was sent again. Scanning all revisions for the same change set therefor allowed for a matching between user name and IP address. This loophole has been closed, however.

the period the bug has been active, some usernames could be mapped. Naturally, the more edits a user did in this period, the more likely is an appearance in this list. For the purpose of this thesis, I screen-scraped the entire table and condensed[32] it to 14,171 unique users.

QUOVA This geo-location web service maps an IP address to a geographic location, see Georeferences.

WIKITRUST Based on Adler et al. [23] an open source online reputation system[33] was set up by the University of California, Santa Cruz, to allow for easy vandalism detection (see Contributions). Given an article ID and a revision ID, the API method *wikimarkup* returns an annotated version of the wikitext of that revision. An annotation consists of a trust value, the revision ID by which the text got introduced into the article as well as the authors user name or IP address, e.g. revision 473029564 of the article *2011-2012 Bahraini uprising*[34]:

Listing 3: Excerpt of the annotated markup for the revision 473029564 of the article *2011-2012 Bahraini uprising*

```
{{#t:7,468889105,Kudzu1}}The
{{#t:7,470041169,Happysailor}}2011-2012
{{#t:8,413989516,Master&Expert}}Bahraini
{{#t:8,427545590,Kudzu1}}uprising, sometimes called the
{{#t:9,455029613,Sitrawi86}}February 14 Revolution
```

All wikitext following an annotation, up to the next one, was written by that author. The web service provider implemented a custom diff algorithm for the attribution of authorship. This was needed to overcome wiki-specific issues and to maximize tracking, e.g. for text that is removed and re-inserted at a later revision.[35]

32 Some usernames have multiple entries as each occurrence of the bug created a unique "evidence". Among those, some have been manually verified and ranked. When multiple entries exist, my algorithm picks the top ranked.

33 http://www.wikitrust.net/vandalism-api (visited on 10/31/2011)

34 http://en.collaborativetrust.com/WikiTrust/RemoteAPI?method=wikimarkup&pageid=30876395&revid=473029564 (visited on 01/23/2012)

35 http://www.wikitrust.net/frequently-asked-questions-faq#TOC-On-text-author-and-origin (visited on 01/24/2012)

4.2 AVAILABLE TOOLS

To process the data from all the data sources, a wide range of software tools are available in the open source community. A simple search for "Wikipedia" on GitHub[36], a source code exchange platform, shows a multitude of small software projects. These come in different programming languages and different feature sets and usually help in downloading articles in batches and extract data from big dump files. Developed by vigilantes and researchers alike, these programs facilitate both data retrieval and processing.

4.2.1 *Toolkits*

A group of openly available software packages[37] qualify as swiss-army knifes for processing and analyzing Database dumps:

PYWIKIPEDIA As the mother of all Python toolkits, the Python wikipedia robot framework[38] offers an extendable set of classes for all MediaWiki entities like a page, a user, revision, etc, and is typically used to write a bot program for automated editing tasks (see Authors).

PYMWDAT Based on PYWIKIPEDIA, this toolkit offers a convenient downloader for all revisions of an article as well as an extensible dump file analyzer with support for filtering and revert detection.

LEVITATION A creative project to turn Database dumps into a Git[39] repository. As a source code management system, Git offers a more space efficient way to store the sequence of revisions of the articles, since it only stores the difference between revisions. Once converted to a repository, moving from revision to revision is much faster than processing the large dump files. Git's diff mechanism together with its *blame* command can be used as an alternative way to attribute authorship to passages of article content.[40]

36 https://github.com/search?q=wikipedia&type=Repositories (visited on 12/12/2011)

37 Although none of these were used for the content analysis of this thesis, their study proved very insightful on how to process Wikipedia's content.

38 http://pywikipediabot.sourceforge.net/ (visited on 01/24/2012)

39 http://git-scm.com/ (visited on 01/24/2012)

40 In fact, git operates on a line level, making the attribution rather coarse. To get blaming functionality on a word level, I patched the source, see my fork at https://github.com/davkal/levitation/commit/5fca0001d26cb67fde6ff9d8a5f2b1414cf7681e (visited on 01/24/2012).

4.2.2 *Analysis projects*

In addition to the toolkits, a handful of research projects exist that process Wikipedia's content. These are purpose built applications that have a much narrower focus but are very skillful in combining and using different data sources such as MediaWiki API, the Toolserver or other Third-party sources/web services:

WIKIPRIDE Python web-application[41] to visualize contributions of groups of editors that registered in the same month.[42]

WIKI TRIP JavaScript application[43], written at SoNet, that uses the MediaWiki API as well as its own Toolserver scripts, to visualize the evolution of a single article over time including: anonymous vs. registered contributors, male vs. female registered users, anonymous edits by country.[44]

4.3 APPLICATION DESIGN

For the analysis of articles I developed an application that could draw data from the different sources and process that data in a timely fashion. Following the impressive WikiTrip application design (see Analysis projects) I decided to build a web application that runs entirely in a web browser.[45]

4.3.1 *Technologies*

As a web application the software heavily relies on HTML version 5, JavaScript (JS), and Cascading Style Sheets (CSS). It uses a range of open-source toolkits and libraries for a variety of purposes:

BOOTSTRAP Twitter's web application toolkit[46] controls the basic layout, the styling of sections and form fields, and the navigation bar at the top.

JQUERY The JS library jQuery[47] is used to asynchronously retrieve data from the various data sources, dynamically in-

41 https://github.com/declerambaul/WikiPride (visited on 12/11/2011)
42 Project website: http://meta.wikimedia.org/wiki/Research:WikiPride (visited on 12/11/2011)
43 https://github.com/volpino/wikipedia-timeline (visited on 12/11/2011)
44 Live demo: http://sonetlab.fbk.eu/wikitrip/ (visited on 12/11/2011)
45 Application development was done mainly using the Google Chrome browser. The application should run in any HTML5-capable browser.
46 http://twitter.github.com/bootstrap/ (visited on 01/25/2012)
47 http://jquery.com/ (visited on 01/25/2012)

sert elements into the layout as well as parsing screen-scraped webpages by making use of its selectors for addressing elements. The AUTOCOMPLETE[48] widget of jQuery's UI library is used to display article suggestions based on what has been entered into the article search field.

UNDERSCORE Underscore[49] is a "utility-belt library" for functional programming in JS. Its *map* and *groupBy* methods are being heavily used in the analysis of content.

BACKBONE Built upon Underscore, Backbone[50] provides a way to structure a JS application. All models, collections, and views in the application are encapsulated by Backbone objects that can communicate with each other via events.

DATEJS Date.js[51] is a library that was being used to parse dates in articles about events.

LZ77 A JS implementation[52] of the LZ77 text compression algorithm[53]. It is used to shrink the size of results when stored in the browser's limited data store[54].

D3.JS Data-driven documents (d3)[55] is a library to visualize big data sets. This application uses d3's box plots[56] to summarize the quantitive distribution of analysis results.

GOOGLE CHART TOOLS Google's[57] chart tools[58] provide a wide range of chart types that I used in the application, including line chart, scatter chart, and motion chart.

4.3.2 *Models*

Treating Wikipedia's website as a system of pages, authors, and revisions, I modeled these as classes of objects using BACKBONE.

When a model is instantiated, it knows where to retrieve the data that will populate its attributes, e.g. a revision collection

48 http://jqueryui.com/demos/autocomplete/ (visited on 01/25/2012)
49 http://documentcloud.github.com/underscore/ (visited on 01/25/2012)
50 http://documentcloud.github.com/backbone/ (visited on 01/25/2012)
51 http://www.datejs.com/ (visited on 01/25/2012)
52 https://github.com/olle/lz77-kit (visited on 01/25/2012)
53 http://en.wikipedia.org/wiki/LZ77_and_LZ78 (visited on 01/25/2012)
54 See http://dev.w3.org/html5/webstorage/ (visited on 01/25/2012) for the limitations.
55 http://mbostock.github.com/d3/ (visited on 01/25/2012)
56 http://mbostock.github.com/d3/ex/box.html (visited on 01/25/2012)
57 Although not open source, some of the charts are developed by the open-source community.
58 http://code.google.com/apis/chart/interactive/docs/index.html (visited on 01/25/2012)

knows that it can get all revisions of an article's history using the query *revisions* from the MediaWiki API. Once the data is retrieved, the attributes of each individual revision object are set, which in turn triggers an event telling a single revision model fetch the annotated wikitext from the WikiTrust API (see Third-party sources/web services). While some events trigger the retrieval of more detailed information, others indicate that a model's data is fully populated and ready for display in the application.

4.3.3 *Views*

The rendering of models is encapsulated in views. Following the publish-subscribe pattern, a view listens to changes in a certain model. When a change event is observed, the view renders itself. For example, the *article view* is rendered multiple times because it draws data from different sources (the page ID is available sooner than, say, the first sentence of the article). Most views, however, rely on several models as they analyze various aspects of an article and then display the results in a chart (see Visualization).

4.3.4 *Article analyzer*

This section describes an application run on a high level. For noteworthy algorithms in the sub-routines, see Algorithms.

When loaded in a browser, the application expects an article title, a category or a template name as input. In the case of an article being entered, a click on "Analyze!" starts the following routine, querying the various Data sources:

1. QUERY INFO is being called to check if the article title is a valid title. A successful query also returns the wikitext of the article's latest revision.
2. A call to PARSE retrieves the HTML version of the latest revision as well as the article's links to other language editions. Both the wikitext and the HTML are then being parsed for location and dates.
3. The SoNet Toolserver script is called to retrieve all authors of the article.
4. For all registered usernames in the author collection, a subroutine is started to locate the users.

5. QUERY REVISIONS is being called to retrieve all revisions (excluding wikitext).
6. For a subset of the revisions, the annotated wikitext is loaded using WikiTrust API.
7. For each language present in the language link collection, the first revision is loaded.
8. All loaded data is analyzed and the results are stored and rendered in the browser.

Depending on the article's number of contributors, this process takes a few minutes and involves 100–500 API calls, most of which are done in parallel. During this process, the browser shows the data whenever it has arrived and been processed.

4.3.5 *Group analyzer*

When a category or template name has been entered, the application switches to *group mode*. On a high level, the group analysis works as follows:

1. Fetch a list of articles by calling QUERY CATEGORYMEMBERS or QUERY EMBEDDEDIN.
2. Run the article analyzer routine on all articles.
3. Compute group results and show them in the browser.

During the group analysis only the article overview will be rendered while all of the article analyzer's steps are being executed. While QUERY EMBEDDEDIN returns a one-dimensional list, QUERY CATEGORYMEMBERS can return subcategories. For a category that has been entered into the input field, the pages of its first-level subcategories will be retrieved as well.

4.4 ALGORITHMS

The individual algorithms described mostly deal with extracting data from the various Data sources. To classify an article as treating an event, the article had to be parsed for a location and a date or date interval. All algorithms will be written in a Python-esque pseudo-code.

4.4.1 *Article requirements*

The presence of a date and a location is a direct requirement for an article to qualify for further analysis when the application is in group mode. All of the following criteria have to be satisfied:

- The article has a location in the form of geographic coordinates or by name, e.g. Cairo.
- The article has a date, e.g. 9 November 1989, or a date interval, e.g. May 2007 - August 2007.
- The article was created after the event started (with a 3 day tolerance) to sort out events that have been scheduled.
- The event (start) date is not before 2002, to make sure that Wikipedia was available as a medium.
- The article does not use certain templates or is part of certain categories, e.g. *Category:Living People*. This has been included for the purpose of filtering out articles that passed the previous tests but are clearly not treating an event.
- At least 25% of authors can be located.

Even when all of the requirements above are met, some articles may not have a full set of results after their analysis. They will however be included in the computations for which results are present, see Hypothesis analysis.

4.4.2 *Date parsing*

The way dates are mentioned in articles are as diverse as the people that write them. And when it comes to date intervals, e.g. May 8–12 2007, even a specialized date parsing library like DATE.JS can only be of limited use. I wrote a custom parser that returns first occurrence of an interval from a text. If no interval was found, the parsing is repeated for a single date.

Some articles embed info box templates (see Templates) that produce annotated markup using the HCARD[59] microformat. The annotations can be easily addressed with CSS selectors[60] and the values conform to the standard for the representation of date

59 http://en.wikipedia.org/wiki/HCard (visited on 01/25/2012)

60 This requires the template to be properly used by the authors, e.g. http://en.wikipedia.org/wiki/Eastern_Front_(World_War_II) (visited on 01/29/2012) where the template *Start date* is used for the start and the end date, producing the two dates with the same annotation: `dtstart`. Another source of error is the order of year, month, and date, e.g. "Municipal Library

and time, International Organization for Standardization (ISO) 8601[61]. In listing 4 the first block tries to parse the microformat annotations. When they are not present, the first info box is checked for a date field. As a last resort, the first sentence and paragraph are scanned for dates to not filter out eligible articles like *2007 Georgian demonstrations*[62]

Listing 4: Date candidates algorithm

```
if 'dtstart' in HTML:   # checking for hcard
    start = datejs.parse(dtstart.text)
    if 'dtend' in HTML:
        end = datejs.parse(dtend.text) # proper interval
    elif 'ongoing' in dtstart.next.text:
        end = today() # ongoing event
    else:
        end = start + 1 # single day event
elif 'date' in templates.infoboxes[0]:
    start, end = custom.parse(templates.infoboxes[0].date) #
        wikitext
if not start:
    start, end = custom.parse(article.first_sentence) # HTML
if not start:
    start, end = custom.parse(article.first_paragraph) #HTML
```

The custom parser then checks for a range of formats including lazy ones like December 14–19, 2008. Over a long iterative process, I identified the following tokens from which to construct the date patterns as regular expressions:

Listing 5: Date tokens

```
ords = ['th', 'st', 'nd', 'rd'];
tokens = {
    M: "({0})".format(MonthNames.join('|')), # months
    D: "(\d{1,2})" + "({0})?".format(ords.join('|')), # day
    Y: "(\d{4})", #year
    T: "(/|-|--|\sto\s|\sand\s)", # interval delimiter
    O: "('*ongoing'*?|'*present'*)", # ongoing event
    F: "From", # ongoing event
    S: "[,\s]*", # whitespace
    P: "\|", # pipe
```

Elevator Coup" happened on 28 January 1908 which was added as `Start date|1908|28|01` to the info box but unpredictably rendered as April 1, 1908, see http://en.wikipedia.org/wiki/Municipal_Library_Elevator_Coup (visited on 01/25/2012).

61 http://en.wikipedia.org/wiki/ISO_8601 (visited on 01/25/2012)

62 The article does not contain an info box and a date is only mentioned in the second sentence: "The demonstrations peaked on November 2, 2007,..." with more dates to follow in the same paragraph. This particular example already shows how idiosyncratic dates can be codified. See http://en.wikipedia.org/wiki/2007_Georgian_demonstrations (visited on 01/25/2012).

```
    A: "([^- --]*)" # other text
};
```

Using these tokens, I produced patterns to match all encountered date formats, e.g. the pattern for December 14–19, 2008 is `MDTDY`, or to capture the even less conform "From 15 October 2011" (meaning the event is ongoing) the pattern is `FDMY`. The patterns are ranked by accuracy so that "12 May 2001 - present" is matched before "May 2001 - present". The coarsest pattern to match is `Y`, a single year (4 digits).

When a date was extracted using my custom parser, its accuracy is also stored as the *date resolution* with one of the following values: *day*, *month*, *year*.

4.4.3 *Location parsing*

Like dates, locations and even coordinates can be codified in numerous ways. Most of the templates for geographic coordinates used in the info boxes produce annotated markup. Although this makes the coordinates machine-readable, I still wrote a custom parser for all the cases where the coordinates marker (`geo`) is not produced.

Some articles do not have coordinates, although they clearly describe an event, e.g. *Maspero demonstrations*[63]. From articles like these, location candidates are scraped and then resolved:

Listing 6: Article's locate algorithm

```
function locate(article):
    if 'geo' in article.HTML: # checking for machine-readable
        coords
        location = custom.parse(article.geo.text)
        return
    else: # look for location candidates
        candidates = []
        # all links from the first info box's location field (
            wikitext)
        candidates.extend(templates.infoboxes[0].location.links
            )
        # all flags in the first info box
        candidates.extend(templates.infoboxes[0].flags) #HTML
        # all links from the first paragraph
        candidates.extend(article.first_paragraph.links) # HTML
    if len(candidates) and article.isMainArticle:
        candidates = candidates[:10]
        until location:
```

63 http://en.wikipedia.org/wiki/Maspero_demonstrations (visited on 01/25/2012)

```
16              location = locate(retrieve(candidates.pop()))
17          return location
```

The above mentioned *Maspero demonstrations* article is exemplary for the candidate list mechanism. Its info box's location field offers three links to articles of a place: Maspiro, Cairo, Egypt. They are checked until an article with coordinates is found, in this case after the second try, Cairo[64].

4.4.4 *Collective authorship*

Most of the authorship processing is being done by SoNet's Toolserver script[65], see Toolserver. The PHP-script directly queries a live copy of the Wikipedia database for all revisions of the requested article. For each revision's author, an edit counter is incremented, and if the author was anonymous, the IP is being resolved to a geographic location. A collection of authors and all resolved locations are then returned in a JSON object.

From the response, my application then creates an author collection. If an author is a registered bot, the author is excluded. From the location list, a second collection is created to manage all author locations for the current article.

The attribution of text passages to authors is done by the web service WikiTrust, see Third-party sources/web services. It returns an annotated markup that can easily be parsed with the following regular expression (JS):

```
/{{#t:\d+,\d+,[^}]*}}/g;
```

Barring bots, all edits are considered relevant, i.e. reverts or blanking[66] are not treated in a special way.

4.4.5 *Locating users*

The location of all anonymous authors has already been determined by SoNet's Toolserver script. For each of the remaining (registered) users, it is first checked if the user is included in the *Poor man's checkuser* list (see Third-party sources/web services). In that case, the username can be resolved to an IP address which, in turn, can be resolved to an accurate location, see IP Look-up.

64 http://en.wikipedia.org/wiki/Cairo (visited on 01/25/2012)

65 https://github.com/volpino/toolserver-scripts/blob/master/\ac{PHP}/api.\ac{PHP} (visited on 01/25/2012)

66 The illegitimate removal of all content of an article, see http://en.wikipedia.org/wiki/Wikipedia:VANDTYPES#Types_of_vandalism (visited on 01/25/2012)

For all remaining authors, their user page is parsed for a location, see Parsing user pages.

4.4.5.1 *IP Look-up*

Resolving the location[67] for a given IP address is a simple call to Quova's IP-lookup API.[68] The limits of the non-commercial license — 2 requests per second and a maximum of 1,000 requests per day — have been overcome by a caching server proxy.

4.4.5.2 *Parsing user pages*

A user page is scanned for possible locations by parsing its HTML content. This is easier than parsing the wikitext, since embedded templates can be inconsistent[69]. The parser looks for links with country names and then checks if they appear in a certain context:

Listing 7: User page location algorithm

```
candidates = []
for link in userpage.links:
    if is_country(link.title):
        candidates.append(link)

patterns = [" comes? from", " am from", "This user is from",
    "This user is in", " lives? in", " currently living in"]

for candidate in candidates:
    context = link.parent
    for pattern in patterns:
        if context.match(pattern):
            country = candidate
            break
return country
```

Plainly parsing for locations and flags, as done in the article's location parsing, yields too many false positives as some user pages are flooded with flags of countries the user professes to have visited. However, looking at the context of where a link appears, helps in parsing prose such as:

> "...and live in a rather small town close to
> `<a title='Austria'>`my country's`</a>` capital".

67 Regarding their accuracy and coverage, their website is rather vague, but coverage is at least 99.8% on a state level, see http://www.quova.com/what/ (visited on 01/25/2012)

68 For the API call, a registered account is needed. This provides a secret key which has to be used to sign each request.

69 See http://en.wikipedia.org/wiki/Category:Nation_of_origin_user_templates (visited on 01/25/2012)

from the user page of *Nightstallion*[70].

4.4.5.3 *Geographic resolution*

As accurate as the coordinates from the IP location services may appear, adding the results of the user page analysis to the set of located authors means having to settle for a country-level resolution.

Both the article location and the user page parsing algorithms are searching for countries. The location for a country is looked up in the application's country list (see COUNTRIES in Wikipedia website). There, each country name is mapped to the geographic coordinates present in its Wikipedia article. For bigger countries, the coordinates refer to the location of the capital, e.g. *United States*[71] while for smaller ones a central point is denominated, e.g. Bahrain[72]. In any case, the coordinates still refer to a point on the globe and, by definition, cannot describe an area. Despite this shortcoming, coordinates are the primary means to reference places in Wikipedia.

As a result, all users that were located by user page parsing, are referenced with coordinates of a country. This has the potential to grossly bias seemingly accurate metrics like the signature distance, described in the next section. For a political analysis, however, it may be less important that all users that have been located in the US by user page parsing, share the same coordinates.

4.4.6 *Signature distance*

Proximity metrics are a means to determine the localness of contributions. Therefore, they play a central part in the application's contribution analysis. In his dissertation, Hardy [24] developed a proximity metric called *signature distance*. This calculates the "average distance between an article and all its contributing authors, weighted by the relative work per author"[73].

Hardy's formula for the signature distance uses basic properties of articles and authors. Using the same notation, let ρ be an author and α be an article. For a sample of articles S let A be

70 http://en.wikipedia.org/wiki/User:Nightstallion (visited on 01/25/2012)

71 http://en.wikipedia.org/wiki/United_States (visited on 01/25/2012)

72 http://en.wikipedia.org/wiki/Bahrain (visited on 01/25/2012)

73 See p. 52 for the complete development of the formula. For clarity, this quotation has been stripped of mathematical symbols.

the set all articles in the sample and P be the set of all located authors in the sample:

$$A = \{\alpha : \alpha \in S\} \qquad P = \{\rho : \rho \in S\}$$

Further, let $\eta(\rho, \alpha)$ denote the contribution(s) of author ρ to article α—an author's "direct work" on an article[33]. Then, $N(\alpha)$ are all contribution(s) to article α:

$$N(\alpha) = \{\eta(\rho, \alpha) : \rho \in P\}$$

Conversely, $P(\alpha)$ are the author(s) who have made contributions to article α:

$$P(\alpha) = \{\rho : \rho \in N(\alpha)\}$$

To calculate the "relative work"[24, p. 52], let $w(\rho, \alpha)$ be relative edit frequency for author ρ on article α:

$$w(\rho, \alpha) = \frac{|\eta(\rho, \alpha)|}{|N(\alpha)|}$$

Let further $\delta(\rho, \alpha)$ be the geodesic distance between author ρ and article α. Then, averaging the distances of authors weighted by their relative work leads to the signature distance $D(\alpha)$:

$$\begin{aligned} D(\alpha) &= \sum_{\forall \rho \in P(\alpha)} \frac{|\eta(\rho, \alpha)| \cdot \delta(\rho, \alpha)}{|N(\alpha)|} \\ &= \sum_{\forall \rho \in P(\alpha)} \big(w(\rho, \alpha) \cdot \delta(\rho, \alpha)\big) \end{aligned} \tag{4.1}$$

Hardy weighted the author distances by their edit frequency and thereby treated all contributions as equal. With more data available, namely the attribution of text to individual authors (see Collective authorship), I can provide two variations of the signature distance metric. The two following metrics only take those revisions into account, that produced the text of the current revision. This is the same text someone reading the article on Wikipedia's website would see, at the moment of the analysis.

When compared to the original signature distance, the following two metrics indicate whether more local or more distant contributions prevailed. In the first variant, edits are still counted as a unit, reflecting the direct work on the article of which parts survived. The second variant looks at how big these parts are in analyzing remaining text lengths of the individual contributions.

4.4.6.1 *Edit survival*

Following the previously used notation, I will develop the *e.surv signature distance* metric. For a given revision, it counts only those contributions that have provided text still present in that revision. Let Π be the set of all revisions in the sample S:

$$\Pi = \{\pi : \pi \in S\}$$

Let $\epsilon(\rho, \alpha, \pi)$ denote the contribution(s) of located author ρ to article α that are still present in revision π. $E(\alpha, \pi)$ are all contribution(s) to article α that are still present in revision π:

$$E(\alpha, \pi) = \{\epsilon(\rho, \alpha, \pi) : \rho \in P\}$$

The weight function $w_{e.surv}$ is then modified to reflect the relative work only of the surviving contributions:

$$w_{e.surv}(\rho, \alpha, \pi) = \frac{|\epsilon(\rho, \alpha, \pi)|}{|E(\alpha, \pi)|}$$

This restriction on edit survival leads to the *e.surv signature distance* for revision π of article α:

$$\begin{aligned} D_{e.surv}(\alpha, \pi) &= \sum_{\forall \rho \in P(\alpha)} \frac{|\epsilon(\rho, \alpha, \pi)| \cdot \delta(\rho, \alpha)}{|E(\alpha, \pi)|} \\ &= \sum_{\forall \rho \in P(\alpha)} \left(w_{e.surv}(\rho, \alpha, \pi) \cdot \delta(\rho, \alpha) \right) \end{aligned} \tag{4.2}$$

Depending on the spatial distribution of an article's authorship and nature of their edits, the values for D and $D_{e.surv}$ can differ significantly. If, over time, distant contributions are replaced by local ones, $D_{e.surv}$ would be much lower than D. If all edits survived, both values would be identical. This very fact allows for the original signature distance to be used as a baseline for the *e.surv index*:

$$I_{e.surv} = \frac{D_{e.surv}(\alpha, \pi)}{D(\alpha)} \cdot 100 \qquad\qquad D(\alpha) \neq 0 \tag{4.3}$$

This normalizes the *e.surv signature distance* to a baseline of 100. Values of $I_{e.surv} < 100$ indicate that local contributions are overrepresented in the current revision, while values of $I_{e.surv} > 100$ indicate an overrepresentation of distant contributions. In the extreme case of $I_{e.surv} = 0$ all surviving[74] contributions came from the article location.

74 This assumes that distant contributions were present in earlier revisions. This also guarantees that $D(\alpha) \neq 0$.

4.4.6.2 *Text survival*

The second variation, the *t.surv signature distance* metric, makes one further adjustment. It weighs the contributions by the length of their text that survived. In other words, for a given revision, it counts how much text can be attributed to which contribution.

To this end, let $\tau(\rho, \alpha, \pi)$ denote the text of contribution(s) of located author ρ to article α that are still present in revision π. $T(\alpha, \pi)$ is all text from contribution(s) to article α that are still present in revision π:

$$T(\alpha, \pi) = \{\tau(\rho, \alpha, \pi) : \rho \in P\}$$

In essence, T is the part of the text in a revision that could be located. The modified weight function then determines an author's share of the located text $w_{t.surv}$:

$$w_{t.surv}(\rho, \alpha, \pi) = \frac{|\tau(\rho, \alpha, \pi)|}{|T(\alpha, \pi)|}$$

This leads to the *t.surv signature distance* for revision π of article α:

$$\begin{aligned} D_{t.surv}(\alpha, \pi) &= \sum_{\forall \rho \in P(\alpha)} \frac{|\tau(\rho, \alpha, \pi)| \cdot \delta(\rho, \alpha)}{|T(\alpha, \pi)|} \\ &= \sum_{\forall \rho \in P(\alpha)} \left(w_{t.surv}(\rho, \alpha, \pi) \cdot \delta(\rho, \alpha) \right) \end{aligned} \tag{4.4}$$

Compared with the original signature distance D, a lower value for $D_{t.surv}$ indicates that text from local contributions fared better than the average contribution. In cases where both values are equal, all text survived of previous revisions survived.[75] When comparing all three values, an order $D > D_{e.surv} > D_{t.surv}$ would suggest that local contributions were more likely to survive and were especially successful as a contribution. A more interesting order would be $D_{e.surv} > D > D_{t.surv}$, suggesting a lot of successful work by distant contributors while the text is still being dominated by text from local contributions.

As in the previous section, I will introduce a new indicator for text survival, the *t.surv index*:

$$I_{t.surv} = \frac{D_{t.surv}(\alpha, \pi)}{D(\alpha)} \cdot 100 \qquad D(\alpha) \neq 0 \tag{4.5}$$

The baseline is again at 100, while values of $I_{e.surv} < 100$ indicate that text from local contributions is overrepresented in the current revision.

75 Or, very unlikely, was scaled in an exactly the same manner.

4.4.6.3 *Implementation of the signature distance*

Given that an article has a location and its authors have been located, their signature distance (equation 4.1) can be calculated using the following implementation:

Listing 8: Signature distance algorithm

```
sd = 0 # signature distance
total = 0 # number of all edits
for author in article.authors:
    if loc in author: # count only authors that have a location
        dist = geodesic(author.loc, article.loc)
        edits = author.count # work is the edit count
        total += edits # postpone relativization to last line
        sd += dist * edits
return sd / total
```

The algorithm in listing 8 relies on the preprocessed edit counts by SoNet's Toolserver script and provides the signature distance only for the latest revision. The computation of the signature distance for all revisions is more expensive. Given at least two differently-located authors, the signature distance changes with each new revision that has a locatable author. That means for each revision, all previously located edits have to be counted. As the computation for revision η_n would have to go over the same edits as the computation for η_{n+1}, the algorithm (see listing 9) uses a technique called *memoization* where previous results are stored within the function for later use.

Listing 9: Signature distance algorithm for all revisions

```
function compute(i, located):
    revision = located[i]
    dist = geodesic(revision.author.loc, article.loc)
    if i == 0:
        return dist
    return (dist + (i - 1) * compute(i - 1, located)) / i;

memoized = memoize(compute)
# only revisions that have a located author
located = revisions.filter_location()
sd = 0
for revision, index in located:
    sd = memoized(index, located)
    revisions.set(sd)
```

The other two variations of the signature distance have been implemented in a similar way.[76]

76 Below, in listing 11, another algorithm based on text survival is presented.

4.4.7 *Localness*

An essential part of the authorship analysis deals with the question whether an author is *local*. If the geographic coordinates are available for an author's location, the geodesic distance can be used to determine the proximity. In this case the users are regarded as local if their geodesic distance to the article location does not exceed a limit. The distance has to be in the lower quartile of all author distances and not be more than 500km.

When the user's location could only be determined by parsing the user page, the user is considered local when the parsed country is the same as the country of the event article.

Listing 10: Localness of an author

```
function is_local(author, article, authorship):
    if author.location:
        if author.location.coordinates:
            # coordinates found via IP lookup
            distance = geodesic(author.location, article.
                location)
            # quartiles of all author distances
            distance_distribution = distance_quartiles(
                authorship)
            # within lower quartile or hard limit of 500 km
            return distance <= min(500,
                distance_distribution.1st_quartile)
        else:
            # country from user pages
            return author.location.country == article.location.
                country
    else:
        return null  # third state, localness is unknown
```

4.5 VISUALIZATION

The algorithmic analyses of the articles are in fact measurements, i.e. multiple series of numbers, that can also be represented in graphical terms. Following the maxim "The purpose of visualization is insight, not pictures."[34, p. 6], the application's diagrams aim to support the presentation of results and to invite exploration. This section will only present the different types of diagrams as they are an integral part of the application. The next chapter, EXPERIMENTS, discusses how the shown results can be interpreted.

4.5.1 *Maps*

To show the location that was extracted from an article, a simple world map is used. A marker shows the position of the location's geographic coordinates.

Figure 3: Article location for *2011–2012 Bahraini uprising*

The author analysis groups the located authors by country. For each country, the number of edits of its authors is counted, resulting in a mapping of country name to edit count. A choropleth map (see figure see figure 4) is used to show how this measure varies over different countries. The darker a country is rendered, the higher is its edit count.

However, this choropleth map only measures the edit counts and does not acknowledge the fact that contributions may disappear by edits in later revisions. Therefore, a second choropleth map (see figure 5) is shown below the first one, displaying text volume based on text survival (see also Contributions). This allows for a direct comparison of contributor countries with a high activity (edit counts) versus contributor countries whose citizens write text that survives the scrutiny of other editors, e.g. Egypt gained intensity (darker green) in the second map.

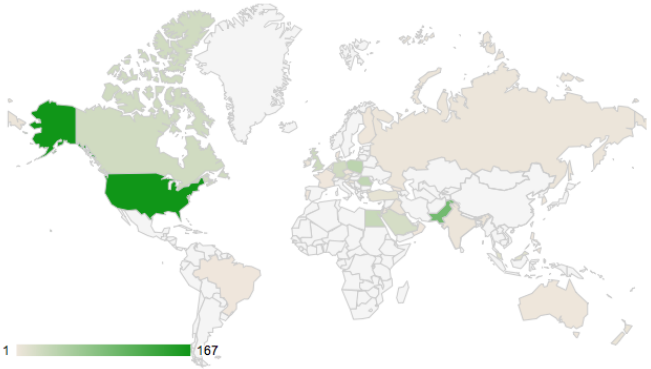


Figure 4: Geographic origins by country for located authors of *2011–2012 Bahraini uprising*

For the second choropleth map, the annotated markup provided by WikiTrust is parsed to extract the authors that intro-

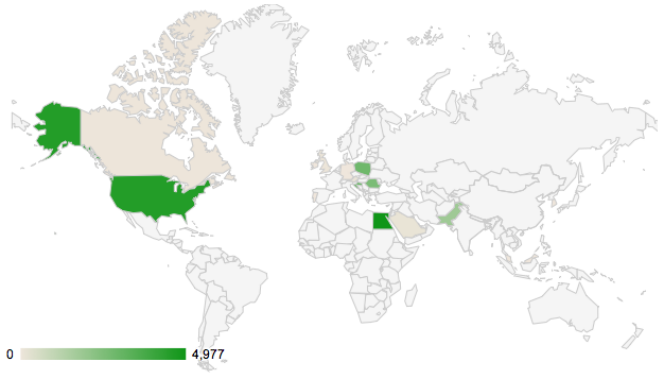


Figure 5: Text survival in revision 471577075 grouped by country for located text of *2011–2012 Bahraini uprising*

duced the text sequences that make up the final text. This subset of revisions is again grouped by the country of its author (if located):

Listing 11: Edit weight for map

```
revision = article.revisions.last
total = revision.length
countries = {} # country name -> [length of sequence, ...]
edits = {} # country name -> proportion of whole text
for author in revision.authors:
    if loc in author: # add only located authors
        if loc.country not in countries:
            countries[country] = []
        countries[country].append(author.text.length)
for country in countries.keys:
    edits[country] = sum(countries[country]) / total
return edits
```

4.5.2 *Line charts, scatter charts and box plots*

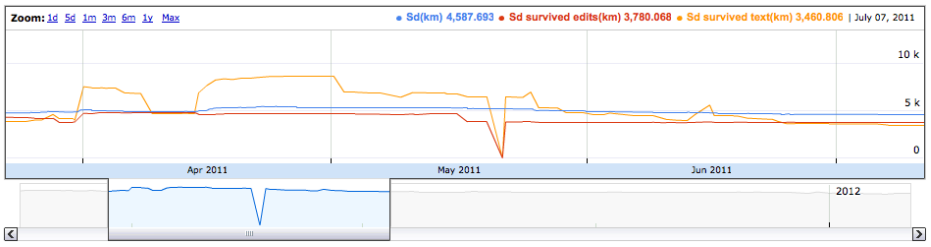


Figure 6: Localness chart for *2011–2012 Bahraini uprising*

As part of the article analysis, a timeline chart shows the evolution of the three proximity metrics over time: the signature distance and the two new variants, *e.surv* and *t.surv*, see figure 6.

For the rendering of the hypotheses test results, also line charts as well as scatter charts and box plots are being used, see chapter RESULTS.

4.5.3 *Motion chart*

The motion chart offers an alternative view on the metrics shown in the choropleth map. This chart type is ideal to follow the change in several indicators over time. Each country is represented by a bubble that, depending on the metric chosen, either moves along an axis or changes its size.[77] The application uses Google's implementation of the motion chart[78]. It is interactive and invites the user to explore the data by allowing the user to freely choose which metric should be represented by which axis.

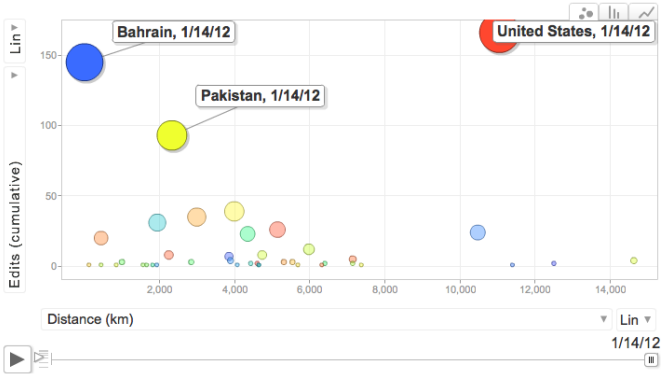


Figure 7: Temporal development of edit counts by country for *2011–2012 Bahraini uprising*

Even clearer than the choropleth map, the motion chart identifies the main contributor countries, e.g. the big blue bubble (Bahrain) and the red one (United States), see figure 7. In the example, changing the metric on the y-axis to show the proportion of located text reveals that contributions from Pakistan have a higher survival rate (the yellow bubble moved up), see figure 8.

4.6 HYPOTHESIS ANALYSIS

The data gathered in the article analysis will be used to test the hypotheses. In addition to the basic requirements (see Article

77 The list of countries has to be limited to the top 10 countries for each measure. Otherwise, the amount of the data would have been too big for a browser application; the worst case, an 11-year-old article with daily contributions from each country, yields more than 1m values per measure.

78 http://code.google.com/apis/chart/interactive/docs/gallery/motionchart.html (visited on 01/25/2012)

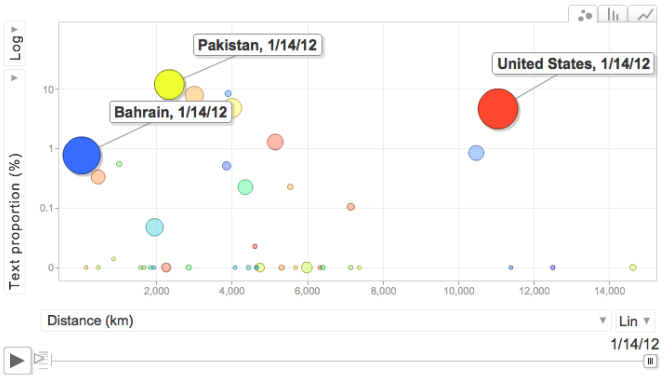


Figure 8: Temporal development of text proportion by country for *2011–2012 Bahraini uprising*

requirements) for each article to qualify, the following sections describe for each hypothesis which data from the content analysis is used in testing support for it.

For hypotheses that rely on a correlation for their support, the Pearson product-moment correlation coefficient is used.

H1: Articles are created shortly after the start of the event.

When the date parsing algorithm finds a date, an article is assumed to treat an event. Moreover, the parser supports different time resolutions. Some events have a precise date-to-date interval, e.g. *2011 Dohuk riots*[79] while others merely happened over the course of a month, e.g. *February 2010 Australian cyberattacks*[80], or even years, e.g. *Oyster Wars*[81].

An article qualifies for this hypothesis if the date parsing resulted in dates with a day or month resolution. The hypothesis is supported when articles were created within a certain time after the event's start date. For an article start date with a day resolution this limit is 7 days; for articles with a month resolution the limit is 30 days.

H2: The more recent an article, the shorter is the delay between the event start and article creation.

The qualification criteria are the same as for H1. For each article the time difference between article creation and event start date is calculated. The hypothesis is supported if there is a linear de-

79 http://en.wikipedia.org/wiki/2011_Dohuk_riots (visited on 01/27/2012)
80 http://en.wikipedia.org/wiki/February_2010_Australian_cyberattacks (visited on 01/27/2012)
81 http://en.wikipedia.org/wiki/Oyster_Wars (visited on 01/29/2012)

pendence between article creation date and the time difference, in the form of a negative correlation coefficient.

H3: Articles are being created first in the English Wikipedia.

Only articles that exist in multiple language editions qualify. Further, the event's location parsing must yield a country where English is not an official language. Then, articles lend support to the hypothesis if they have been created first in the English Wikipedia.

The language codes used to identify Wikipedia's language editions "mostly correspond to the language codes defined by ISO 639-1".[35] This allows to test whether a language edition corresponds to one of a country's official languages (see OFFICIAL LANGUAGES in Wikipedia website).

H4: Articles about political events are created by local authors.

Articles qualify if their creator has been located. An article lends support to the hypothesis if the creator is considered local, see Localness.

H5: In the beginning most contributions are anonymous.

Articles qualify that have at least 10 revisions (excluding bots) in the time interval after creation. The interval for event articles with day resolution is 7 days; the limit for articles with a month resolution is 30 days. An article lends support to the hypothesis when the number of anonymous contribution is bigger than the number of contributions done by registered users.

H6: In the beginning most contributions are local.

Articles qualify when they have at least 10 revisions (where the author was located, excluding bots) in the time interval after creation. An article lends support to the hypothesis when the number of local contribution is bigger than the number of distant, i.e. non-local, contributions (see Localness).

H7: Later, the share of anonymous contributions decreases over time.

Articles qualify that have revisions in at least 6 different months. Revisions are grouped by month. The hypothesis is supported

if there is a linear dependence between the article age and the share of anonymous contributions per month, in the form of a negative correlation coefficient.

H8: Later, the share of local contributions decreases over time.

Articles qualify that have revisions in at least 6 different months. Revisions are grouped by month. The hypothesis is supported if there is a linear dependence between the article age and the share of local contributions per month(see Localness), in the form of a negative correlation coefficient.

H9: Local contributions are more likely to survive.

Articles qualify that have at least 10 located revisions. Local contributions are more likely to survive when, on average, the *e.surv* signature distance is smaller than the original signature distance. Using the *e.surv index*, the hypothesis is supported if for all revisions the mean *e.surv index* is smaller than 100.

H10: Text from local contributions is more likely to survive.

Articles qualify that have at least 10 located revisions. Text from local contributions is more likely to survive when, on average, the *t.surv* signature distance is smaller than the original signature distance. Using the *t.surv index*, the hypothesis is supported if for all revisions the mean *e.surv index* is smaller than 100.

4.7 POSSIBLE ENHANCEMENTS

The previously mentioned algorithms can be improved in various ways. In this section I'll suggest starting points for enhancements.

4.7.1 *Edit relevance*

Improvements regarding edit relevance aim to filter revisions that do not seem to change the meaning of an article. Barring revisions coming from bots, all revisions are being treated as relevant. It may be worth investigating whether revisions marked as minor edits[82] should be counted as full contributions. Also, detecting vandalism could help in filtering for relevant edits, al-

82 See Revision history.

though it can be argued that vandalism is just another form of contribution, albeit a rather drastic one.

4.7.2 *User page parsing*

Parsing the user pages is a problem of information extraction. Currently, the pages are scanned for a fixed number of patterns, e.g. "lives in…". The list of patterns could be extended and tested against user pages. One candidate for this are the WikiProject[83] info boxes which registered users can put on their user pages. From a box like *Template:WikiProject New York City*[84] an affiliation with New York City could be inferred.

The current algorithm is scanning for links to Wikipedia and then checks if they appear in the right context. For prose that has not been formatted with links, e.g. "I am from Warsaw." a web service like WIKIPEDIAMINER[85] could be used.

4.7.3 *Geographic profiling*

Lieberman and Lin [26] describes an algorithm to determine the location of a registered author based on the pages the author contributes to. An implementation would load a user's contribution (MediaWiki API query USERCONTRIBS), fetch these pages and scan for coordinates, construct a convex hull around them and then determine the center.

83 http://en.wikipedia.org/wiki/Wikipedia:WikiProject (visited on 01/27/2012)

84 http://en.wikipedia.org/wiki/Template:WikiProject_New_York_City (visited on 01/27/2012)

85 Website: http://wikipedia-miner.cms.waikato.ac.nz/ (visited on 01/27/2012), Example "I am from Warsaw.": http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/?source=I+am+from+Warsaw. (visited on 01/27/2012)

5 EXPERIMENTS

The application described in APPARATUS will now be fed with articles. This chapter will demonstrate results for individual articles and give pointers on how they can be interpreted. For a quantitative analysis testing the hypotheses on a set of articles, see the next chapter RESULTS.

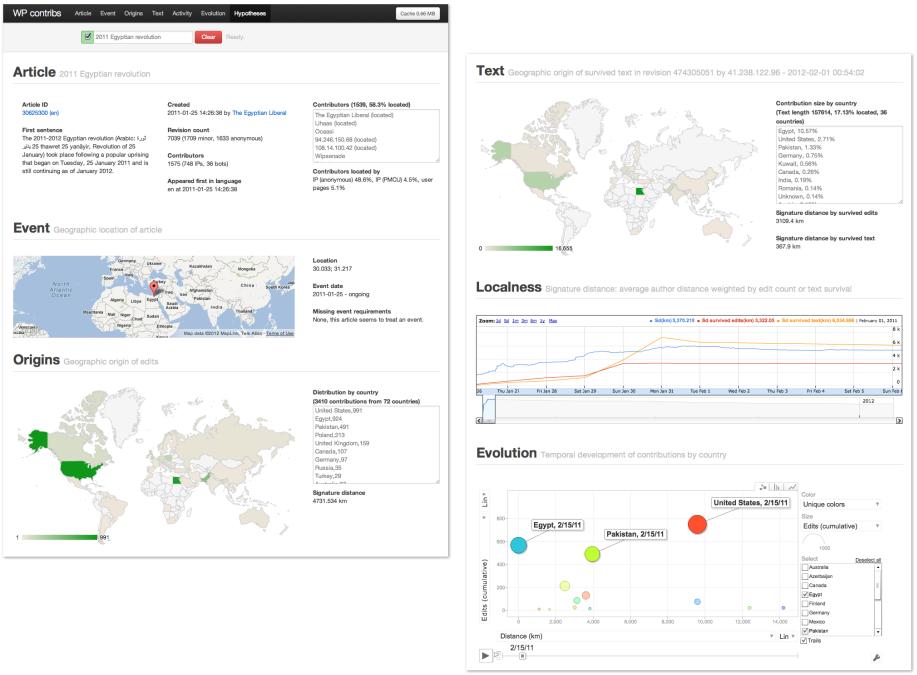


Figure 9: Sample run for article *2011 Egyptian revolution*

5.1 EXAMPLE ARTICLES

The experimental runs will be done on the following articles:

EGYPT-REV The article *2011 Egyptian revolution*[1] details the protests leading to the overthrow of the government.

LIBYA-WAR The *Libyan civil war*[2] refers to the country-wide rebellion that turned into a civil war during the Arab Spring.

1 http://en.wikipedia.org/wiki/2011_Egyptian_revolution (visited on 02/04/2012)

2 http://en.wikipedia.org/wiki/Libyan_civil_war (visited on 02/04/2012)

BAHRAIN-UP The *2011–2012 Bahraini uprising*[3] refers to the events surrounding a series of demonstrations in the Gulf state during the Arab Spring.

The application analyzes the articles' contributions and presents the results in a report with different sections, see figure 9.

5.2 ANALYSIS

Some metrics of the three runs are shown in table 1. The rows *Located contributors* show the effectiveness of the georeferencing algorithm using different data sources. Using the combined techniques of the Poor Man's Checkuser database and the parsing of user pages increases the share of located contributors by more than 20%.

5.2.1 *Proximity metrics*

It is also worth comparing the proximity metrics in the lower two groups of rows in table 1. The values from the *cumulative* group are based on the edit counts of all contributions. For the second group of rows, only contributions were counted that provided text still *present in the latest revision*. Over time, all three articles collected contributions from a number of countries. However, during the collective editing process, up to two thirds of them were removed, as can be observed for BAHRAIN-UP.

The signature distance variant *e.surv* can hint whether local contributions are more likely to survive. A lower *e.surv* value compared to the original signature distance suggests that contributions from distant countries did not survive as well as local contributions. For EGYPT-REV, the *e.surv* distance is 28% lower than the original signature distance, suggesting that distant contributions were more likely to be removed or replaced by local ones.

More strikingly, comparing the *t.surv* distance to the other distance metrics reveals how misleading metrics based on edit count can be. The *t.surv* distance weighs an author's distance to the article's location by the relative amount of text that survived. A low *t.surv* distance means that texts from local contributions dominate the article—despite the other metrics suggesting numerous contributions from more distant countries. EGYPT-REV shows this behavior where the signature distance shrank by 43%

3 http://en.wikipedia.org/wiki/2011-2012_Bahraini_uprising (visited on 02/04/2012)

Table 1: Example articles

	EGYPT-REV	LIBYA-WAR	BAHRAIN-UP
Created	2011-01-25	2011-02-01	2011-02-15
First in language	English	n/a[c]	Swedish
Latest revision	474830650	472144068	474621605
Size (kB)	152.6	174.5	120.7
Revisions	7,045	9,534	1,804
Contributors[a]	1,543	2,093	407
– anonymous	751	928	138
Located contributors (%)	58.4	54.3	44.5
– anonymous (%)	48.7	44.3	33.9
– by PMCU[b] (%)	4.5	4.0	3.7
– by user page (%)	5.2	6.0	6.9
Cumulative			
Unique countries of origin	71	80	40
Signature distance (km)	4,731.5	6,877.1	4,152.0
Present in latest revision			
Unique countries of origin	37	48	16
Signature dist. e.surv (km)	3,684.2	6788.8	3,657.4
Signature dist. t.surv (km)	2,705.0	6797.6	2,190.3
e.surv index	77.9	98.7	88.1
t.surv index	57.2	98.8	52.8
Located text (%)	23.8	9.5	45.2

[a] Contributors are all authors excluding bots.

[b] Poor man's checkuser

[c] The result was Galician. However, the article LIBYA-WAR was merged into the article *Libya* which was created in 2005, thereby preceding the civil war articles in other language editions.

to 2,705km when weighed by survival of text from located contributions. This last qualification is important. If only few of the surviving contributions can be located, the *t.surv* distance becomes less representative—for the example articles only 23.8% of the text passages could be located.

5.3 DISTRIBUTION OF EDIT COUNTS VS TEXT SURVIVAL

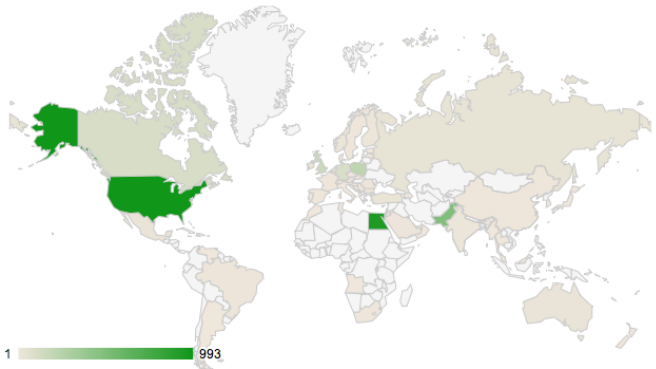


Figure 10: Distribution of origins by cumulative edit count for EGYPT-REV.

The change in the distribution of countries can also be observed in the choropleth maps. Figure 10 shows the map for the cumulative edit count for EGYPT-REV. The number of edits per country is used to determine the intensity of the shade, e.g. 993 edits could be attributed to the United States.

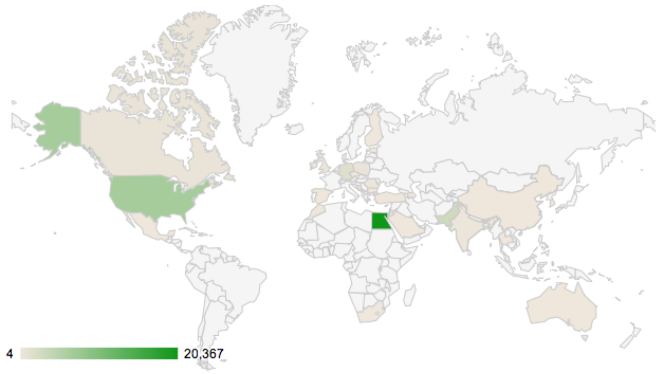
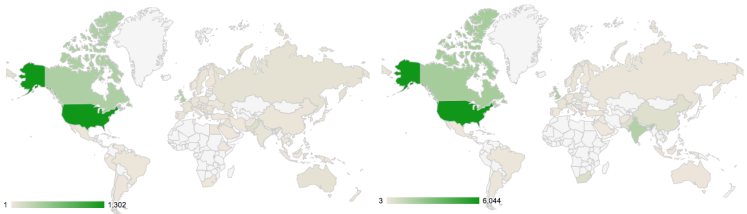


Figure 11: Distribution of origins by text survival for EGYPT-REV.

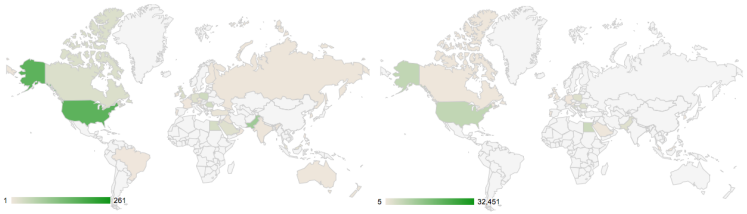
Figure 11 shows the map for text survival; the shader value is the number of characters originating from a country, e.g. of all located survived text, 20,367 characters could be attributed to Egypt.

Differences in shade when comparing both maps can be revealing. Both the United States and Egypt have a similar cumu-



(a) By edit count (b) By text survival

Figure 12: Distribution of origins for LIBYA-WAR.



(a) By edit count (b) By text survival

Figure 13: Distribution of origins for BAHRAIN-UP.

lative edit count, 993 and 925, respectively. However, the United States pale in the text survival map as the located text seems to be clearly dominated by Egypt, with 6,628 and 20,367 characters, respectively. This suggests the article was equally "worked on" by authors from both countries, but contributions from one country were more likely to prevail. It should be noted again, as only 23.8% of the text could be located, no firm conclusions can be drawn.

This effect can also be noticed for BAHRAIN-UP where the United States pale in the text survival map, see figures 13a and 13b.[4] Although 21% of located contributions originated in the United States, only 12% of located text can be attributed to the US; Bahrain contributed 58% of located text.

The maps for LIBYA-WAR are less suggestive, as the United States dominate both distributions, see figure 12. This is also reflected by the similarity of the different signature distance metrics: the values for *e.surv* and *t.surv* differ only by 1% from the original signature distance. Remarkably, less than 1% of contributions could be attributed to Libya, hinting at massive lack of participation.

4 Although it should be shaded darkest, tiny Bahrain cannot be seen in the application's choropleth maps.

5.4 TEMPORAL DEVELOPMENT

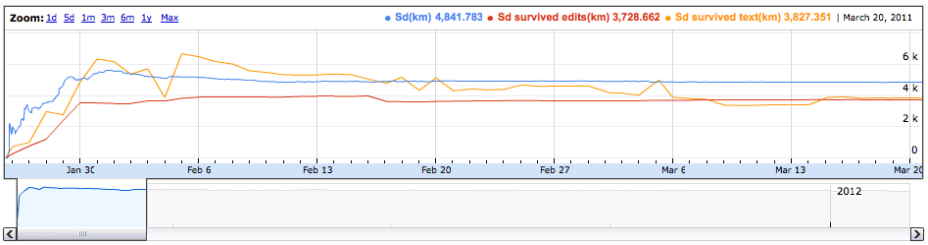


Figure 14: Timeline chart for the signature distance of EGYPT-REV.

In the application's section *Localness*, a timeline chart displays the evolution of the signature distance and its two variations over time, see figure 14. There, the blue line is the original signature distance. An $x - y$—point on that line reflects the average distance y of all located contributions before the date x weighed by their relative work.

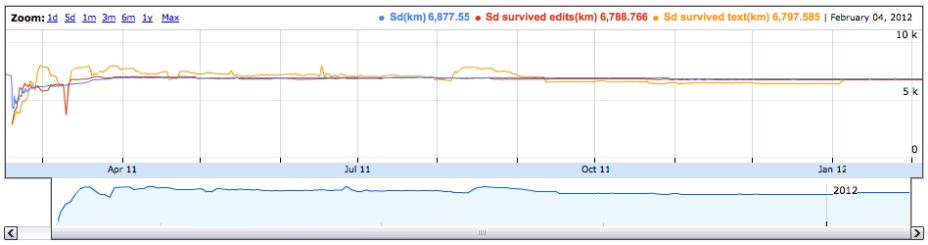


Figure 15: Timeline chart for the signature distance of LIBYA-WAR.

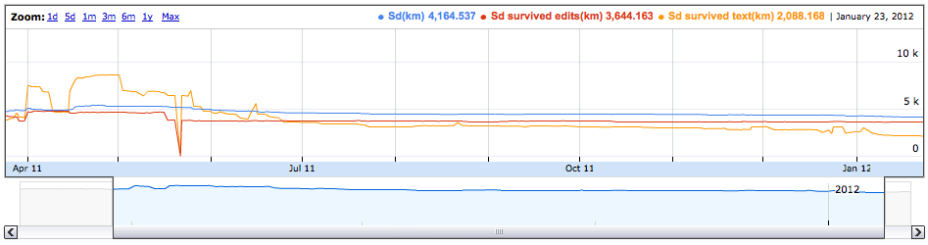


Figure 16: Timeline chart for the signature distance of BAHRAIN-UP.

The orange line is the *t.surv* signature distance, that weighs the contributions by text survival. When the orange line is above the blue line, text from distant contributions is overrepresented[5] in the current revision. Whenever the orange or red (*e.surv*) lines cross the blue line from above, this overrepresentation changes from distant to local. Both EGYPT-REV and BAHRAIN-UP show this behavior, see figure 14 and see figure 16. For LIBYA-WAR the lines

5 The baseline is the original signature distance (blue line).

are always close together, as suggested by their similar signature metrics, see figure 15

A second visualization of the temporal development is in the section *Evolution*. There, a motion chart presents the evolution of key metrics over time. For this, all located contributors were grouped by country and are represented by a bubble. The cumulative edit count, the number of survived edits, or the share of survived text can all be plotted against the distance to the article's event location or against each other.

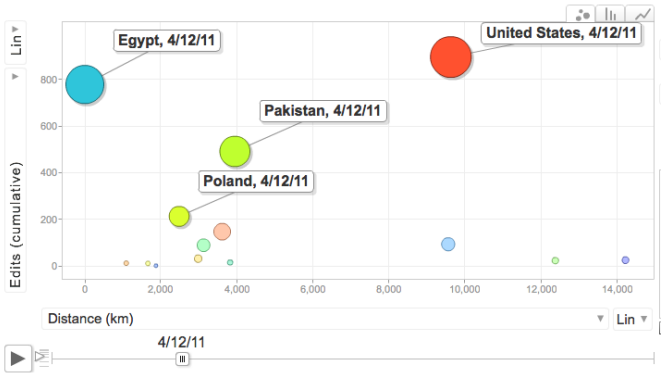


Figure 17: Evolution of cumulative edit count vs. distance for EGYPT-REV.

By default, the cumulative edit count (y-axis) is plotted against the distance (x-axis), see figure 17. An increase in edit count is represented in two ways: a bubble's vertical location, as well as the bubble's size. When the motion chart is played, countries with very active contributors move up faster and grow in size. The locality of a bubble and its size make it easy to spot the biggest contributors and their spatial relationship to the article's location, e.g. a big bubble in the upper right represents a distant country with lots of contributions.

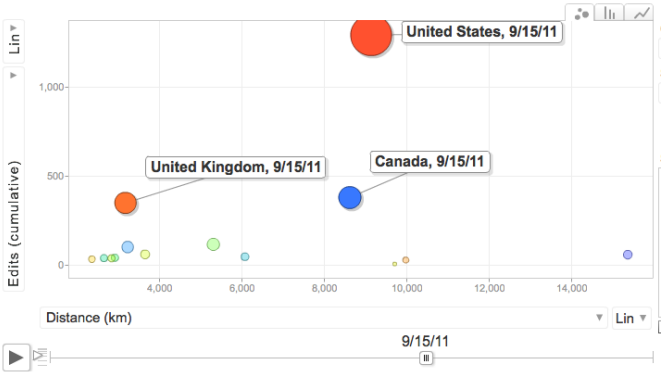


Figure 18: Evolution of cumulative edit count vs. distance for LIBYA-WAR.

When comparing the charts for the cumulative edit count vs. distance of all three articles, see figures 17, 18, and 19, LIBYA-WAR stands out. Libya, the country treated in the article, is not among the top ten contributors, see figure 18. Although already noted in the previous section Distribution of edit counts vs text survival, the lack of contributions becomes more apparent in the motion chart since it is less dependent on a country's size for visual recognition.

The motion chart for BAHRAIN-UP also benefits from this charts' independence of country size, showing Bahrain's comfortable lead in cumulative edits, see figure 19.

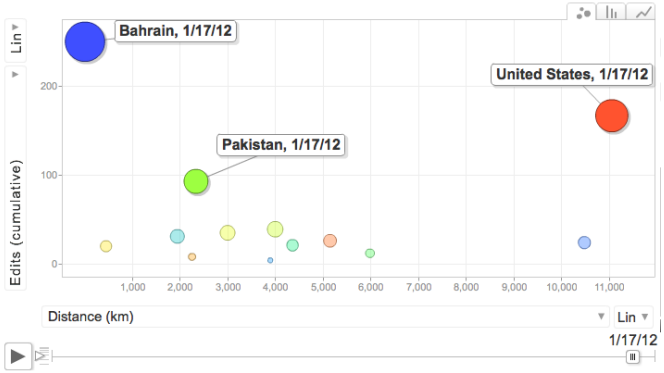


Figure 19: Evolution of cumulative edit count vs. distance for BAHRAIN-UP.

When changing the y-axis to text survival[6], the bubbles shift to reflect the countries' share of the text in the latest revision. Here, the biggest move downwards suggests that a lot of a country's contributions did not survive.

Another interesting plot can be achieved by leaving the y-axis on text survival while setting the x-axis to cumulative edits, see figure 20. This plot reflects the effectiveness of edits. Bubbles in the upper right represent hard earned text survival by lots of edits. Conversely, a bubble in the lower right suggests contributors of a country produce a lot of edits that do not survive.

5.5 HYPOTHESES ANALYSIS

In a final experiment, I tested the three articles for their hypotheses support. The results are in table 2.

6 The text survival metric is best viewed in a logarithmic scale because the chart axis is scaled according to a country's maximum share. A country's share is usually between 1 and 15% but can potentially be 100 % at some point in time—usually in an early stage of the article, or later as a result of vandalism.

Table 2: Hypothesis support of example articles
• article lends support; ○ article does not lend support

HYPOTHESIS	EGYPT-REV	LIBYA-WAR	BAHRAIN-UP
H1 Created instantly	• 1d	• −14d[e]	• 1d
H2 Recent art. sooner[a]	n/a	n/a	n/a
H3 First in English	•	○ Galician[f]	○ Swedish
H4 Creator was local	•	n/a[b]	n/a[b]
H5 Early: more anon.	○ 23% anon.	○ 25% anon.	○ 19% anon.
H6 Early: more local[c]	○ 27% local	○ all distant	○ 8% local
H7 Later: less anon.[d]	○ R = 0.0	○ R = 0.0	○ R = 0.0
H8 Later: less local	○ R = 0.0	○ R = 0.0	○ R = 0.28
H9 Local edit survival	• $\overline{\text{I}_{\text{e.surv}}} = 77.6$	• $\overline{\text{I}_{\text{e.surv}}} = 99.4$	• $\overline{\text{I}_{\text{e.surv}}} = 83.8$
H10 Local text survival	• $\overline{\text{I}_{\text{e.surv}}} = 71.5$	○ $\overline{\text{I}_{\text{t.surv}}} = 102.1$	• $\overline{\text{I}_{\text{e.surv}}} = 87.2$

[a] Only applicable to a group of articles.

[b] Unable to locate creator.

[c] Share of located contributions.

[d] No correlation found.

[e] A negative date is the result of an article being merged with another article treating earlier events.

[f] In the Galician Wikipedia, LIBYA-WAR was merged into the article *Libya* which was created in 2005, thereby preceding the civil war articles in other language editions.

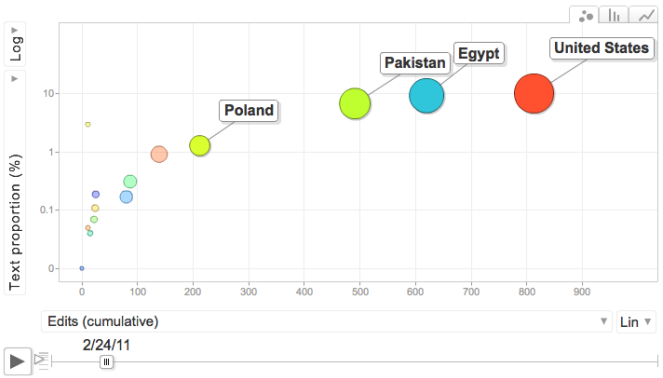


Figure 20: Evolution of text survival vs cumulative edit count for EGYPT-REV.

Although all three articles could be tested properly, each test for hypothesis support seems to be accompanied by its own set of problems. The test for H1 is susceptible to article merges. When an article is not deemed important enough to exist on its own, it will be merged into a related article by the community.[7] The merged article is replaced by a redirect that forwards visitors to the parent article. When this parent article is older than the event date, it results in a negative delay, e.g. −14 days for LIBYA-WAR.

The merging problem also seems to affect H3 for LIBYA-WAR. There, the event article was merged into the article for Libya[8] which was created in 2005. The hypothesis test orders the articles from different language editions by creation date. Therefore, articles merged into old parent articles are in effect shifting their creation date into the past, allowing them to rank first.

Support for H5 and H6 seems to be consistently absent. Neither local nor anonymous contributions account for the majority of contributions. Support for H7 and H8 seem to be equally hard to come by. Only BAHRAIN-UP shows a slight correlation between article age and share of local contributions. Unfortunately, a positive correlation coefficient indicates a rising share, see figure 21.

Figure 21: Share of local users (1.0 = 100%) for BAHRAIN-UP.

7 Reasons for a merge include "significant content overlap" and lack of context covered by a broader topic. See http://en.wikipedia.org/wiki/Wikipedia:Merging (visited on 02/04/2012)

8 http://gl.wikipedia.org/wiki/Libia (visited on 02/04/2012)

Regarding H9 and H10, the articles EGYPT-REV and BAHRAIN-UP would lend some support. EGYPT-REV and BAHRAIN-UP displayed a strong overrepresentation of local text ($I_{t.surv} = 57.2$ and $I_{t.surv} = 52.8$, respectively) in the latest revision, see table 1. However, the mean over all revisions was $\overline{I_{e.surv}} = 71.5$ and $\overline{I_{e.surv}} = 87.2$, respectively, see table 2. This suggests the overrepresentation of local text was not always present, as the charts of the two indexes attest, see figures 22 and 23. As expected, LIBYA-WAR does not show such a shift, see figure 24.

The next chapter RESULTS conducts a quantitative analysis on a set of articles and tests them for hypothesis support.

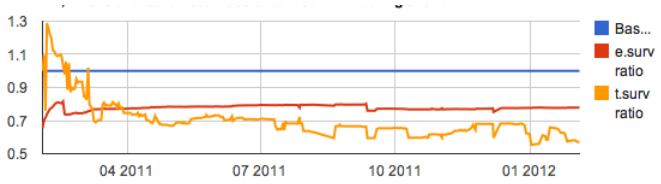


Figure 22: *e.surv* and *t.surv* indexes (1.0 = 100%) for EGYPT-REV.

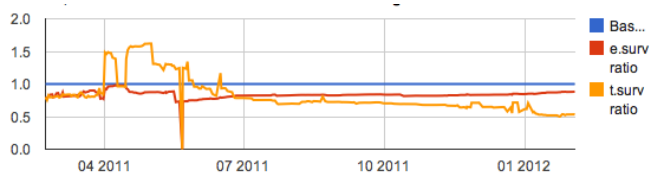


Figure 23: *e.surv* and *t.surv* indexes (1.0 = 100%) for BAHRAIN-UP.

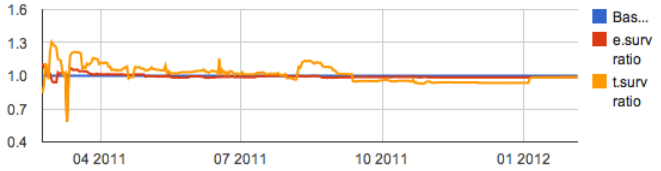


Figure 24: *e.surv* and *t.surv* indexes (1.0 = 100%) for LIBYA-WAR.

Part III

RESULTS

6 RESULTS

This chapter present the quantitative analysis of a set of articles of political events from the English Wikipedia. First, I will describe the Data set and the difficulties in picking the "right" category of articles. Then, I will explain the results of the application's analysis of the data set and their relevance for the hypotheses.

6.1 DATA SET

Unfortunately, *political events* are a rather vague concept judging by the contents of the Wikipedia category of the same name[1]. It includes only one page at the top-level, *No Berlusconi day*. It contains numerous subcategories, however, ranging from *Political riots* and *Protests* to *Political party assemblies*.

In contrast to the categories system, templates offer a way to identify more homogenous groups of articles.[2] Articles that embed a template, e.g. *Infobox civil conflict*, can be identified and grouped, resulting in a set of articles that are very likely to treat civil conflicts.

In order to find a suitable set of articles, I will combine both approaches. From the English Wikipedia I will handpick categories and templates and combine their articles into one data set.

Since this process involves a manual selection, articles are not selected at random. As a result, statistics derived from the analysis cannot be representative of all articles treating political events—let alone the complete corpus of the English Wikipedia. On the bright side, the handpicked categories contain articles that were themselves manually chosen by the community to be included those categories.

1 http://en.wikipedia.org/wiki/Category:Political_events (visited on 02/01/2012)

2 See also Templates.

6.1.1 *Selected categories and templates*

Wikipedia's category system suffers from several forms of over-categorization.[3] Some category sub-trees are too fine-grained and have numerous levels with only a small number articles. When moving along the tree in search for articles of political conflicts, the potential to go astray is rather high, e.g. the category *Protest*[4] has the subcategory *Protest songs*. There is also considerable overlap between sub-trees, e.g. *Protests by country* and *Protests by year*. As a compromise, the category traversal will only yield the pages of a given category and all pages of its subcategories.

Protests and its subcategories are exhaustive and produce a handful of candidates. To these I will add the category *Arab Spring* and the template *Infobox civil conflict*[5] to assert that the articles detailed in EXPERIMENTS are part of the data set. The final data set CONFLICTSRAW combines articles from the following categories and templates:

- Protests by year
- Protests by country
- 2000s riots by year
- 2010s riots by year
- Arab Spring
- Infobox civil conflict (template)

Other notable categories that I investigated but not included, are *21st-century conflicts*[6], *Revolutions categories*[7], as well as the templates *Infobox military conflict*[8] and *Infobox election*[9].

3 For an overview of issues see http://en.wikipedia.org/wiki/Wikipedia:OCAT (visited on 02/04/2012)

4 http://en.wikipedia.org/wiki/Category:Protests (visited on 02/04/2012)

5 This template is also the preferred infobox template of the Occupy movement. http://en.wikipedia.org/wiki/Occupy_movement (visited on 02/04/2012)

6 Only 10% of the 833 articles would have qualified. http://en.wikipedia.org/wiki/Category:21st-century_conflicts (visited on 02/04/2012)

7 This category and its subcategories contain too many articles about people. http://en.wikipedia.org/wiki/Category:Revolutions_by_country (visited on 02/04/2012)

8 A huge set of 10290 articles that is dominated by historic battles. Less than 6% of these articles would have qualified. http://en.wikipedia.org/wiki/Template:Infobox_military_conflict (visited on 02/04/2012)

9 These articles are usually created long before the election date, as elections tend to be scheduled. http://en.wikipedia.org/wiki/Template:Infobox_election (visited on 02/04/2012)

6.1.2 *Data set preparation*

When articles for CONFLICTSRAW are retrieved, they are only added if they are not already in the set. Therefore, all articles in the set are unique. The titles of all articles in CONFLICTSRAW are listed in the appendix, see Data set: CONFLICTSRAW.

Then, the all articles in CONFLICTSRAW are tested whether they fulfill the Article requirements for an analysis. Of the 742 articles in CONFLICTSRAW, 334 passed this first test. The main reason for rejection was an event date before 2002. For the distribution of failed requirements see figure 25. All articles that fulfill the requirements are included in the data set CONFLICTS, see Data set: CONFLICTS in the appendix for a list of titles.

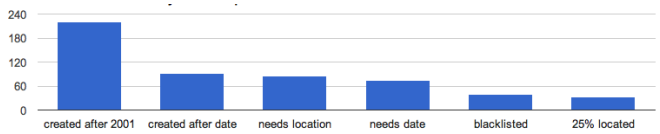


Figure 25: Distribution of failed article requirements

6.2 CHARACTERISTICS OF THE SET

Each article in the set has a location attribute. Therefore the spatial distribution of these articles can be plotted on a map.

Figure 26: Spatial distribution of articles in data set CONFLICTS.

Table 3: Metrics (mean) for articles in CONFLICTS

Age	693.6d
Size (kB)	26.3
Revisions	303.5
Contributors[a]	81.8
– anonymous	37.3
Located contributors (%)	49.8
– anonymous (%)	40.0
– by PMCU (%)	5.2
– by user page (%)	4.6
Cumulative	
Unique countries of origin	8.8
Signature distance (km)	4,271.3
Present in latest revision	
Unique countries of origin	5.4
Signature dist. e.surv (km)	4,465.1
Signature dist. t.surv (km)	4374.6
e.surv index	105.6
t.surv index	106.3
Located text (%)	33.9

During the data set analysis I measured the similar metrics as in the EXPERIMENTS. Their means across all articles are shown in table 3.

For each of the hypotheses the articles in the set are tested whether they qualify, i.e. have enough data for an analysis, see figure 27. The requirements for each hypothesis are defined in Hypothesis analysis. The tests to qualify for H5 and H6 are the most restrictive as they require a certain number of revisions in the early interval of the event. Also, the test for H3 to the requirement is quite restrictive as it rules out events from countries with English as an official language. The qualification for H4 requires the article creator to be located, that seems to be also rarely the case.

6.3 H1 – H4: ARTICLE CREATION

On average, event articles are created 19.7 days after the start date of the event. For articles that only have a month resolution,

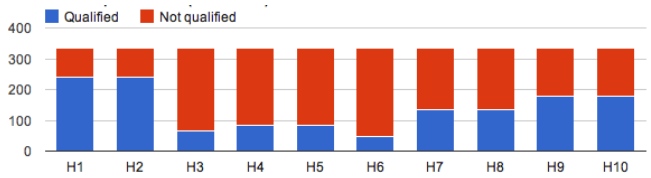


Figure 27: Article qualification for each hypothesis

the average is 223.1 days. These average delays are longer than the stated limit of 7 days and 30 days, respectively. However, the frequency distribution of delays shows a high number of articles publish with only a short delay, see figure 28.

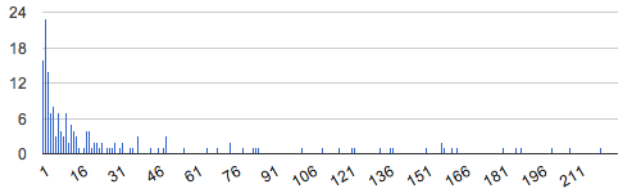


Figure 28: Distribution of creation delays in days

There is only a slight correlation between recentness and creation delay ($R = 0.1$), suggesting the opposite of H1. The delay after which an article is created is more likely to be bigger the more recent an article is. For a scatter plot of creation delay against creation date, see figure 29. The data for H1 and H2 has been cleared of outliers.[10]

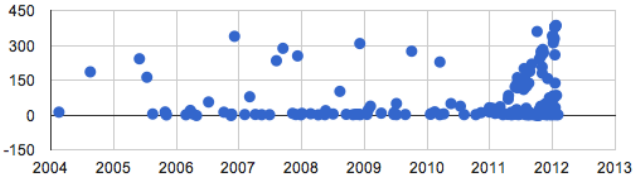


Figure 29: Creation delays over time

Of the 62 articles qualified for H3, 64% (25) were created first in the English Wikipedia. The second most popular language in the set to first publish an article in, was Arabic, see figure 30 For H3, only articles qualified that treat events in countries that do not have English as an official language. The fact that the majority of these articles is published in English, suggests that article creators may prefer a global reach over sharing information with fellow citizen.

10 The lowest value still within 1.5 interquartile range (IQR) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.

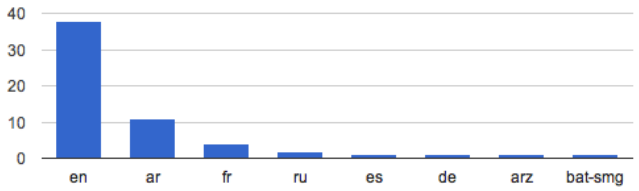


Figure 30: Language editions chosen for first article

H4, namely, that articles about an event in a country are created by a resident, cannot be confirmed. The creator of an article could only be located for 79 articles (11%) in the set. Among those, 89% came from a different country, see figure 31.

Figure 31: Creator localness

6.4 H5 – H8: PARTICIPATION

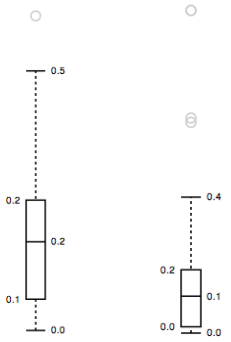
Of the 71 articles that qualified for H5, the average share of anonymous contributions in the beginning is 20%. Therefore, H5 cannot be confirmed. The highest share in the set was 50%, see figure 32a.

For H6, even fewer articles qualified (n = 42). The mean share of local contributions in the beginning is 10%, thereby rejecting H6. See figure 32b for the distribution.

Over time, the divisions anonymous vs registered and local vs distant do not seem to matter much as the mean correlation coefficient is 0 for both cases, see figure 33. Therefore, H7 and H8 cannot be confirmed (n = 119).

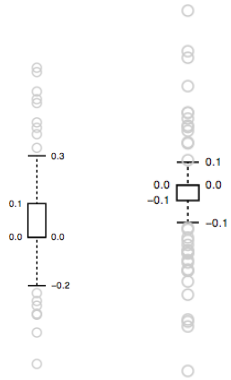
6.5 H9 – H10: TEXT SURVIVAL

Across the set, there is only a slight tendency for local content to be overrepresented. The mean *e.surv index* is 97; the mean *t.surv index* is 98. These index values a only slightly smaller than the baseline of 100. Therefore, H9 and H10 cannot be confirmed. The data for H9 and H10 has been cleared of outliers using a



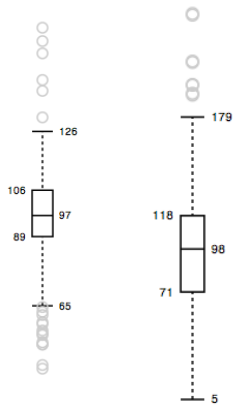
(a) Anon. (b) Locals

Figure 32: Distribution of contributor ratios in early revisions



(a) Anon. (b) Local

Figure 33: Distribution of correlation coefficients for the share of anonymous and local authors, respectively



(a) $\overline{I_{e.surv}}$ (b) $\overline{I_{t.surv}}$

Figure 34: Distribution of means of localness indexes

very broad range.[11] The box plot shows a number of outliers[12], suggesting that there are articles that are dominated by local text, see figure 34.

11 The lowest value still within 4 IQR of the lower quartile, and the highest datum still within 4 IQR of the upper quartile.

12 These outliers are defined by the box plot. The lowest value is still within 1.5 IQR of the lower quartile, and the highest datum still within 4 IQR of the upper quartile.

7 CONCLUSION

In this thesis, I set out to determine whether articles of political events were written by the people most affected, i.e. in the event's proximity. To investigate, I developed a web application that georeferenced contributions to those articles and analyzed them. For the analysis, I presented a locating algorithm that can also locate a part of the registered users.

Then, I combined the georeferenced contributions with a source for text attribution. This allowed an article's content to be split into passages written by different located authors. In effect, this put a distance on a passage of text written by a located author.

The attribution of text to authors also allowed me to address the issue of text survival. With this in mind, I devised two variants of a proximity metric, the signature distance, that are sensitive to edit survival. The final variant, *t.surv*, calculated the signature distance of the text that survived to the current revision.

For three exemplary articles I showed how theses metrics can be used to analyze the localness of an article. Then, I applied the same method on a set of articles about political conflicts. Across the set, the results were mostly inconclusive. Weak support was found for Wikipedia's role as a news medium while studying article creation.

But, as the experiments have shown, there is merit in an individual article analysis. The difference in participation was starkest between the articles of the Bahraini uprising and the Libyan civil war. Contributions from Libya to the article treating its current event, were almost absent. On the other hand, tiny Bahrain contributed the bulk of the content to its article, truly writing its own history.

7.1 FURTHER RESEARCH

Future research can build on the application and tweak it from the technology side, as suggested in Possible enhancements. There is also the possibility to analyze the data already gathered in more refined ways.

The same hypotheses may be applicable to other types of articles than political ones. The key requirements are that they have

a location attribute and a time interval. This is easily fulfilled by disaster articles, e.g. Fukushima Daiichi nuclear disaster[1].

Although *location* is central to more abstract concepts like *Culture*[2] these subjects clearly defy being attributed with *a* location. Nevertheless, an analysis of the spatial distribution of contributors could be interesting.

Finally, articles covering the same topic across various language editions could be analyzed to further investigate issues like language barrier—locals contributing only in their language—and information arbitrage as suggested by Adar, Skinner, and Weld [28].

1 http://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster (visited on 10/31/2011)

2 http://en.wikipedia.org/wiki/Culture (visited on 10/31/2011)

Part IV

APPENDIX

A

APPENDIX

A.1 DATA SET: CONFLICTSRAW

742 articles:
1025 rally to safeguard Taiwan
15 October 2011 global protests
1740 Batavia massacre
1886 Belfast riots
1936 Syrian general strike
1937 peasant strike in Poland
1956 Georgian demonstrations
1957 Alexandra Bus Boycott
1958 Grozny riots
1965 Yerevan demonstrations
1968 Polish political crisis
1968 Red Square demonstration
1971 Łódź strikes
1973 Uruguayan general strike
1978 Georgian demonstrations
1981 warning strike in Poland
1987 Mecca incident
1988 Polish strikes
1989 Sukhumi riots
1990s uprising in Bahrain
1992 Coalisland riots
1993 Cherbourg incident
1993 Russian constitutional crisis
1997 rebellion in Albania
2000 Cochabamba protests
2000 Ramallah lynching
2001 Bradford riots
2001 England riots
2001 Harehills riot
2001 Jos riots
2001 Oldham race riots
2002 Gujarat violence
2003 Maldives civil unrest
2003 Phnom Penh riots
2003 Port of Oakland dock protest
2004 Al-Qamishli riots
2004 Haitian coup d'état
2004 Palm Island death in custody
2004 Redfern riots
2004 unrest in Kosovo
2005 Ahvaz unrest
2005 Alexandria riot
2005 Belize unrest
2005 Birmingham race riots
2005 Cronulla riots
2005 Macquarie Fields riots
2005 Maldives civil unrest
2005 Planoise Forum fire
2005 Toledo Riot
2005 anti-Japanese demonstrations
2005 civil unrest in France
2006 Aligarh Riots
2006 Basel Hooligan Incident
2006 Brussels riots
2006 Dalit protests in Maharashtra
2006 Dublin riots
2006 Ferentari riot
2006 G20 ministerial meeting
2006 Indian anti-reservation protests
2006 Islamist demonstration outside the Embassy of Denmark in London
2006 Israeli reserve soldiers' protest
2006 Nuku'alofa riots
2006 Oaxaca protests
2006 United States immigration reform protests
2006 civil unrest in San Salvador Atenco
2006 democracy movement in Nepal
2006 protests in Hungary
2006 student protests in Chile
2006 youth protests in France
2006–2008 Lebanese political protests
2007 A.S. Roma–Manchester United F.C. conflict
2007 Bersih rally
2007 Burmese anti-government protests
2007 Catania football violence
2007 Georgian demonstrations
2007 HINDRAF rally
2007 Iranian petrol rationing riots
2007 Karachi riots
2007 Koidu-Sefadu protest
2007 MacArthur Park rallies
2007 Macau labour protest
2007 Macau transfer of sovereignty anniversary protest
2007 Russian protests
2007 Venezuelan demonstrations
2007 West Bengal food riots
2007 civil unrest in Villiers-le-Bel
2007–09 university protests in France
2007–2008 Kenyan crisis
2008 Armenian presidential election protests
2008 Cameroonian anti-government protests
2008 Congo football riots
2008 Egyptian general strike
2008 Greek riots
2008 Icelandic lorry driver protests
2008 Jos riots
2008 Longnan riot
2008 Otago NORML protests
2008 Podgorica protest
2008 Shenzhen anti-police riot
2008 Sichuan riots
2008 South Korean candlelight vigil
2008 Tibetan unrest
2008 UEFA Cup Final riots
2008 US beef protest in South Korea
2008 Weng'an riot
2008 global rice shortage
2008 riot in Mongolia
2008 student protests in Chile
2008 unrest in Bolivia
2008–2009 Anti-Israel riots in Norway
2008–2010 Thai political crisis
2009 Albina, Suriname riots
2009 Ashura protests
2009 California college tuition hike protests
2009 G-20 London summit protests
2009 Georgian demonstrations
2009 Gojra riots
2009 Guinea protest
2009 Icelandic financial crisis protests
2009 Luquan protest
2009 Malagasy political crisis
2009 May Day protests
2009 Mexico prison riot
2009 Nigerian sectarian violence
2009 Peruvian political crisis
2009 Riga riot
2009 Tamil diaspora protests
2009 Tamil diaspora protests in Canada
2009 Upton Park riot
2009 student protests in Austria
2009 student protests in Croatia
2009–2010 Iranian election protests
2010 Air Untuk Rakyat rally
2010 Canada anti-prorogation protests
2010 Catalan autonomy protest
2010 G-20 Toronto summit protests
2010 Hong Kong new year march
2010 Jos riots
2010 Karachi riots
2010 Kyrgyzstani revolution
2010 Macau labour protest
2010 Macau transfer of sovereignty anniversary protest
2010 Santa Cruz riots
2010 South Kyrgyzstan riots
2010 Suzhou workers riot
2010 Thai military crackdown
2010 Thai political protests
2010 Tibetan language protest
2010 UK student protests
2010 Xinfa aluminum plant protest
2010 student protest in Dublin
2010–2011 Greek protests
2010–2011 Ivorian crisis
2010–2011 Senegalese protests
2010–2012 Algerian protests
2011 Albanian opposition demonstrations
2011 Algerian self-immolations
2011 Anshun incident
2011 Anti-budget demonstration
2011 Armenian protests
2011 Azerbaijani protests
2011 Bahrain Grand Prix
2011 Belarusian protests
2011 Bolivian protests
2011 Burkinabè protests
2011 Chaozhou riot
2011 Chilean protests
2011 Chinese pro-democracy protests
2011 Damascus bombings
2011 Djiboutian protests
2011 Dohuk riots
2011 Egyptian revolution
2011 England riots
2011 Georgian protests
2011 Haimen protest
2011 Indian anti-corruption movement
2011 Indiana legislative walkouts
2011 Iranian protests
2011 Iraqi protests
2011 Israeli border demonstrations
2011 Israeli social justice protests
2011 Kurdish protests in Iraq
2011 Kurdish protests in Turkey
2011 Lebanese protests
2011 Libyan rape allegations

2011 London anti-cuts protest
2011 Macedonian protests
2011 Magallanes protests
2011 Malawian protests
2011 Maldivian protests
2011 Mangystau riots
2011 Mexican protests
2011 Nanchang mass suicide protest
2011 Northern Ireland riots
2011 Oakland general strike
2011 Omani protests
2011 Palestinian protests
2011 Portuguese protests
2011 Qianxi riot
2011 Rome demonstration
2011 Shanghai riot
2011 Spanish protests
2011 Sri Lanka worker protests
2011 Sudanese protests
2011 Sulaymaniyah protests
2011 United Kingdom anti-austerity protests
2011 United States public employee protests
2011 Vancouver Stanley Cup riot
2011 Western Saharan protests
2011 Wisconsin protests
2011 Yunnan protest
2011 Zengcheng riot
2011 Zhili riot
2011 Zhongshan riot
2011 attack on the British Embassy in Iran
2011 attack on the Israeli Embassy in Egypt
2011 land acquisition protests in Uttar Pradesh
2011–2012 Bahraini uprising
2011–2012 Damascus clashes
2011–2012 Daraa Province Clashes
2011–2012 Idlib Governorate clashes
2011–2012 Jordanian protests
2011–2012 Kuwaiti protests
2011–2012 Libyan factional fighting
2011–2012 Mauritanian protests
2011–2012 Moroccan protests
2011–2012 Saudi Arabian protests
2011–2012 Syrian uprising
2011–2012 Yemeni uprising
2012 Romanian protests
228 Incident
27th G8 summit
314 Taipei protest
3rd Summit of the Americas
4th Summit of the Americas
517 Protest
70 Million Steps Against Coups
8888 Uprising
Abdelaziz Bouteflika
Aboriginal Day of Action
Abu Musab Abdel Wadoud
Abuja bus crash riots
Aftermath of the 2011 Libyan civil war
Aftermath of the Bronze Night
Ahmad Al-Shahhat
Ahmed Rifaat (judge)
Akali movement
Al Jazeera
Al-Islah
Al-Nidaa Brigade
Al-Wasat Party
Aleppo protests
Ali Abdulemam
Ali Belhadj
Alliance of Yemeni Tribes
Alliance of Youth Movements
American Majority
Anti-Gaddafi forces
Anti-Roma protests in Bulgaria 2011
Anti-nuclear protests in the United States
April 2009 Thai political unrest
April 6 Youth Movement
Arab League Monitors in Syria
Arab Spring
Arizona State Prison Complex – Lewis
Arthur Uther Pendragon
Ash-shab yurid isqat an-nizam
Assassination of Benazir Bhutto
August 31, 1982 demonstrations in Poland
Avenue Habib Bourguiba
Awn Shawkat Al-Khasawneh
Ayat Al-Qurmezi
BART Police shooting of Oscar Grant
Bahrain Aid Flotilla
Bahrain Independent Commission of Inquiry
Bahrain: Shouting in the Dark
Baltic Way
Barbie Liberation Organization
Batepá massacre
Battle of Brightlingsea
Battle of Douma
Battle of Rastan
Battle of Sana'a (2011)
Battle of Ta'izz
Battle of Zabadani
Battle of Zinjibar
Battle of the Bogside
Belgrade anti-gay riot
Benton Harbor riots
Berkeley Marine Corps Recruiting Center protests
Berkeley oak grove controversy
Bersih 2.0 rally
Best Bakery case
Birmingham campaign
Bisho massacre
Black Friday (2005)
Black January
Black May (1992)
Black Procession
Black Spring (Kabylie)
Bloody Sunday (1969)
Bloody Thursday (2011)
Bolivian miners' protest of 2007
Brian Haw
Bridge of Flowers (event)
Brigade 93
Bristol riots
Bronze Night
Bureau of Indian Affairs building takeover
CU project controversy
California Proposition 8
Camp Trans
Cape Town peace march
Caracazo
Caricature during the 2011 Libyan civil war
Casualties of the 2011 Libyan civil war
Casualties of the 2011–2012 Bahraini uprising
Center for Socialist Studies
Chartism
Children's Crusade (civil rights)
Cholera Riots
Cincinnati riots of 2001
Coalition of Socialist Forces
Coalition of the Youth of the Revolution
Columbia University protests of 1968
Constituent Assembly of Tunisia
Copenhagen December Riot
Cottage cheese boycott
Cow head protests
Crimean anti-NATO protests of 2006
David D. Pearce
Dazhou protests of 2007
Death of Ahmed Jaber al-Qattan
Death of Ali Abdulhadi Mushaima
Death of Ali Jawad al-Sheikh
Death of Fadhel Al-Matrook
Death of Hamza Ali Al-Khateeb
Death of Ian Tomlinson
Death of Khaled Mohamed Saeed
Death of Muammar Gaddafi
Death of Seyed Ali Mousavi
December 2001 riots in Argentina
December 2005 protest for democracy in Hong Kong
December 2011 Syrian–Turkish border clash
Defiance Campaign
Democratic Party (Libya)
Dissenters March
Dockum Drug Store sit-in
Domestic responses to the 2011 Egyptian revolution
Dongzhou protests
Drumcree conflict
Dushanbe riots
EDSA III
EDSA Revolution of 2001
EU-Latin America summit of 2004 protest activity
Early 2012 Hong Kong protests
East L.A. walkouts
Egyptian Tank Man
Egyptian constitutional referendum, 2011
Egyptian parliamentary election, 2011–2012
Egyptian presidential election, 2012
El-Ghad Party
Elizabeth Simbiwa Sogbo-Tortu
Emad Effat
Eman al-Nafjan
End Water Poverty
Environmental direct action in the United Kingdom
Fadi Elsalameen
Farhad Khoiee-Abbasi
Fathers 4 Justice protests
February 14 Youth Coalition
February 2010 Australian cyberattacks
February 28, 2004 hand-in-hand rally
Ferzat Jarban
Firjan
First Quarter Storm
Football riots Široki Brijeg-Sarajevo
Ford (HM Prison)
Fort Qualls
Frack Off
Free Egyptians Party
Free Libyan Air Force
Free Libyana
Free Syrian Army
Free speech in the media during the 2011 Libyan civil war
Freedom of assembly in Russia
Fuel protests in the United Kingdom
Gaddafi's response to 2011 Libyan protests
General People's Congress (Yemen)
Ghad El-Thawra Party
Glasgow school closures protest, 2009
Great American Boycott
Guantánamo Bay hunger strikes
Guinea-Bissau riot, 2007
Gulbarg Society massacre
Gwangju Democratization Movement
Hadi al-Mahdi
Hama
Harold Sturtevant
Hatoon al-Fassi
Higher Political Reform Commission
Hilo Massacre
History of the Muslim Brotherhood in Egypt
Homes before Roads
Homs Bombardment (2012)
Homs airbase ambush
Homs massacre

Houphouët-Boigny Arena stampede
Human Price of Freedom and Justice
Human chain against nuclear plant in Turkey
Human rights in Egypt under the Supreme Council of the Armed Forces
Human rights reports on 2011–2012 Bahraini uprising
Human rights violations in the 2011 Libyan civil war
Humanitarian situation during the 2011 Libyan civil war
Hunger marches
Impact of the Arab Spring
International reaction to the Gaza War
International reactions to the 2011 Egyptian revolution
International reactions to the 2011 Libyan civil war
International reactions to the 2011 military intervention in Libya
International reactions to the 2011–2012 Bahraini uprising
International reactions to the 2011–2012 Syrian uprising
International reactions to the 2011–2012 Yemeni uprising
International reactions to the Arab Spring
International reactions to the Tunisian revolution
International reactions to the death of Muammar Gaddafi
International recognition of the Syrian National Council
Iranian Revolution
Isa Qassim
Jadaliyya
January 19th incident
January 2012 al-Midan bombing
January 27, 2007 anti-war protest
Jastrzębie-Zdrój 1980 strikes
Jeltoqsan
John Sinclair Freedom Rally
July 2009 French riots
July 2009 Ürümqi riots
June 1976 protests
Kaohsiung Incident
Kefaya
Khaled al-Johani
Khalida Toumi
Kosovo is Serbia protest 2008
Law enforcement and the Occupy movement
Legion Thoria
Liberty Street Protest
Libya Contact Group
Libyan Freedom and Democracy Campaign
Libyan League for Human Rights
Libyan civil war
List of Tea Party protests, 2009
List of Tea Party protests, 2010
List of protests in the United Kingdom
List of riots in Leeds
List of road protests in the UK and Ireland
Local Coordination Committees of Syria
London Conference on Libya
Louisa Hanoune
Lublin 1980 strikes
Mabahith
Macassar Village Land Occupation
Magarha
Malari incident
Maldon Grain Riots
Manal al-Sharif
Manama Paper
Manningham riot
March 19, 2008 anti-war protest
March 9, 1991 protest
Marouf al-Bakhit
Maryam Alkhawaja
Maspero demonstrations
May 1968 in France
May 1998 riots of Indonesia
May 2007 RCTV protests
Maya Evans
Mexicans Without Borders
Mid 2011 Telangana protests
Mihalis Filopoulos
Milan Rai
Million Fax on Washington
Million Voices Against Corruption, President Chen Must Go
Million Woman March
Minneapolis Teamsters Strike of 1934
Miss America protest
Mohamed Bouazizi
Mohamed Larbi Zitout
Montebello High School flag flipping incident
Mourad Medelci
Movement for Justice and Development in Syria
Muammar Gaddafi's response to the 2011 Libyan civil war
Muslim Brotherhood
Muslim Brotherhood in Egypt
N2 Gateway occupations
Nabeel Rajab
Naji Fateel
National Association for Change
National Coordination Committee for Democratic Change
National Hunger March, 1932
National Liberation Army (Libya)
National Transitional Council
Navi Mumbai Holi riots
Network of Free Ulema – Libya
New Castle Correctional Facility
No Berlusconi day
November 15, 2008 anti-Proposition 8 protests
Novocherkassk massacre
Occupy Ashland
Occupy Atlanta
Occupy Austin
Occupy Baltimore
Occupy Berlin
Occupy Boston
Occupy Buffalo
Occupy Canada
Occupy Charlotte
Occupy Charlottesville
Occupy Chicago
Occupy Cincinnati
Occupy Cork
Occupy Dallas
Occupy Dame Street
Occupy Edinburgh
Occupy Eugene
Occupy Glasgow
Occupy Houston
Occupy Las Vegas
Occupy London
Occupy Los Angeles
Occupy Melbourne
Occupy Nashville
Occupy Nigeria
Occupy Oakland
Occupy Philadelphia
Occupy Pittsburgh
Occupy Portland
Occupy Providence
Occupy Redwood City
Occupy Reykjavik
Occupy Sacramento
Occupy Salem
Occupy San Diego
Occupy San Francisco
Occupy San Jose
Occupy Seattle
Occupy St. Louis
Occupy Starbucks
Occupy Sydney
Occupy Texas State
Occupy Toronto
Occupy Wall Street
Occupy Windsor
Occupy movement
October 2011 Jabal al-Zawiya clashes
Operation Egypt
Operation Tunisia
Oscar Olivera
Osh riots (1990)
Overthrow of Slobodan Milošević
Oxford Oath
Oyster Wars
Pacification of Wujek
Pearl Roundabout
Peninsula Shield Force
People's Guard (Libya)
People's Olympiad
Persiankiwi
Phosphorite War
Polish 1970 protests
Political crisis in Thailand (2005–2006)
Poll Tax Riots
Poor People's Campaign
Poor People's World Cup
Port Said Stadium clashes
Port of Tacoma protests, March 2007
Poznań 1956 protests
Pro Government Rallies in Syria
Protest in South Africa
Protests against Proposition 8 supporters
Protests against SOPA and PIPA
Protests against the 2011 military intervention in Libya
Protests during the EU summit in Gothenburg 2001
Protests following the 2011 Russian elections
Protests of Wukan
Protests regarding 2008 South Ossetia war
Public opinion of the 2006 Thai coup d'état
Purple Rain Protest
Putin must go
Qorvis
Queers Undermining Israeli Terrorism
Rainbow Nation Peace Ritual
Rand Rebellion
Reactions to Occupy Wall Street
Reform League
Refugees of the 2011–2012 Syrian uprising
Republic Protests
Respect for America's Fallen Heroes Act
Revolutionary Socialists (Egypt)
Rif Dimashq blockades
River Run Riot
Road protest in the United Kingdom
Romanian Revolution of 1989
Ruby Ridge
Russian March
Saad Eddin Ibrahim
Samir Rifai
Sampit conflict
San Bernardino punk riot
San Francisco newspaper strike of 1994
San Luis Obispo Mardi Gras controversy
Saudi Arabian municipal elections, 2011
Save Zimbabwe
Saïd Sadi
Science Is Vital rally
Scientology and the Internet
Seattle Mardi Gras Riots
Selma to Montgomery marches
Sensitive urban zone
September 15, 2007 anti-war protest

September 2009 Xinjiang unrest
September 24, 2005 anti-war protest
Shabeeha
Shaoguan incident
Sharad Pawar slapping incident
Sharpeville massacre
Shishou incident
Siege of Baniyas
Siege of Daraa
Siege of Deir ez-Zor
Siege of Hama
Siege of Homs
Siege of Idlib
Siege of Jisr al-Shughur
Siege of Latakia
Siege of Rastan and Talbiseh
Siege of Talkalakh
SlutWalk
Social situation in the French suburbs
Socialist Popular Alliance Party
Society for Development and Change
Sokwanele
Solidarity Day march
Soweto uprising
Stephen Gough
Strategy-31
Summer 1981 hunger demonstrations in Poland
Swampy
Symphony Way Pavement Dwellers
Syrian National Council
Syrian Observatory for Human Rights
Syrian Revolution General Commission
Syrian conflict peace proposals
SyrianFreePress
Ta-pa-ni Incident
Tahrir Square
Tahrir Square, Alexandria
Taking Back South Africa!
Tanjung Priok massacre
Taxpayer March on Washington
Tea Party Express
Tea Party movement
Tea Party of Nevada
Tea Party protests
Tent State University
Thammasat University massacre
The Propagandists
The Revolution Continues Alliance
Tiananmen Square protests of 1989
Tiger (organisation)
Timeline of 2011 England riots
Timeline of the 2009 Iranian election protests
Timeline of the 2011 Egyptian revolution
Timeline of the 2011 Egyptian revolution under Hosni Mubarak's rule
Timeline of the 2011 Egyptian revolution under Supreme Council of the Armed Forces
Timeline of the 2011 Libyan civil war and military intervention (16 August – 23 October)
Timeline of the 2011 Libyan civil war and military intervention (19 March – May)
Timeline of the 2011 Libyan civil war and military intervention (June – 15 August)
Timeline of the 2011 Libyan civil war before military intervention
Timeline of the 2011 military intervention in Libya
Timeline of the 2011–2012 Bahraini uprising
Timeline of the 2011–2012 Saudi Arabian protests
Timeline of the 2011–2012 Saudi Arabian protests (January–April 2011)
Timeline of the 2011–2012 Saudi Arabian protests (May–December 2011)
Timeline of the 2011–2012 Saudi Arabian protests (from January 2012)
Timeline of the 2011–2012 Syrian uprising
Timeline of the 2011–2012 Syrian uprising (January–April 2011)
Timeline of the 2011–2012 Syrian uprising (May–August 2011)
Timeline of the 2011–2012 Syrian uprising (September–December 2011)
Timeline of the 2011–2012 Syrian uprising (from January 2012)
Timeline of the 2011–2012 Yemeni uprising
Timeline of the 2011–2012 Yemeni uprising (23 September – December 2011)
Timeline of the 2011–2012 Yemeni uprising (3 June – 22 September 2011)
Timeline of the 2011–2012 Yemeni uprising (January – 2 June 2011)
Timeline of the 2011–2012 Yemeni uprising (from January 2012)
Timeline of the Libyan civil war
Torture during the 2011–2012 Bahraini uprising
Trial of Lex Wotton
Trials and judicial hearings following the 2011 Egyptian revolution
Tribute FM
Tripoli
Tripoli protests and clashes (February 2011)
Tunisian revolution
Turn Your Back on Bush
UAE Five
UK Uncut
UK miners' strike (1984–1985)
US domestic reactions to the 2011 military intervention in Libya
USAbilAraby
Ulster Workers' Council strike
Umma Islamic Party
United Nations General Assembly Resolution 65/265
United Nations Security Council Resolution 1970
United Nations Security Council Resolution 1973
United Nations Security Council Resolution 2009
United Nations Security Council Resolution 2016
United States involvement in the 2011 Libyan civil war
University of California, Berkeley
University on the Square
Valdez Blockade
Venezuelan general strike of 2002–2003
Viomak
Vita Cortex sit-in
Voice of Free Libya
Walk for Life West Coast
Washington A16, 2000
Washington for Jesus
Whites Only Scholarship
Winooski 44
Women to drive movement
Workers Democratic Party
World Trade Organization Ministerial Conference of 1999 protest activity
World Youth Day 2011
Wrocław football riot 2003
Xenophobia in South Africa
Yemeni presidential election, 2012
Yizhou riots
Yosberides
Zakariya Rashid Hassan al-Ashiri
Zenga Zenga
Zhejiang solar panel plant protest
Zohra Drif

A.2 DATA SET: CONFLICTS

280 articles:
1025 rally to safeguard Taiwan
2002 Gujarat violence
2003 Maldives civil unrest
2003 Phnom Penh riots
2003 Port of Oakland dock protest
2004 Al-Qamishli riots
2004 Haitian coup d'état
2004 Palm Island death in custody
2004 Redfern riots
2004 unrest in Kosovo
2005 Ahvaz unrest
2005 Birmingham race riots
2005 Cronulla riots
2005 Macquarie Fields riots
2005 Maldives civil unrest
2005 Toledo Riot
2005 civil unrest in France
2006 Aligarh Riots
2006 Basel Hooligan Incident
2006 Brussels riots
2006 Dublin riots
2006 Ferentari riot
2006 G20 ministerial meeting
2006 Nuku'alofa riots
2006 civil unrest in San Salvador Atenco
2006 democracy movement in Nepal
2006–2008 Lebanese political protests
2007 A.S. Roma–Manchester United F.C. conflict
2007 Catania football violence
2007 Georgian demonstrations
2007 Iranian petrol rationing riots
2007 Karachi riots
2007 Venezuelan demonstrations
2007 West Bengal food riots
2007 civil unrest in Villiers-le-Bel
2007–2008 Kenyan crisis
2008 Congo football riots
2008 Egyptian general strike
2008 Greek riots
2008 Jos riots
2008 Longnan riot
2008 Shenzhen anti-police riot
2008 Sichuan riots
2008 Tibetan unrest
2008 UEFA Cup Final riots
2008 Weng'an riot
2008–2009 Anti-Israel riots in Norway
2009 Albina, Suriname riots
2009 Gojra riots
2009 Icelandic financial crisis protests
2009 Malagasy political crisis
2009 Mexico prison riot
2009 Nigerian sectarian violence
2009 Riga riot
2009 Upton Park riot
2010 Jos riots
2010 Karachi riots
2010 Kyrgyzstani revolution
2010 South Kyrgyzstan riots
2010 Thai military crackdown
2010 UK student protests
2010–2011 Ivorian crisis
2011 Anshun incident
2011 Chaozhou riot
2011 Damascus bombings
2011 Dohuk riots
2011 England riots
2011 Egyptian revolution
2011 Haimen protest
2011 Indiana legislative walkouts
2011 Israeli border demonstrations
2011 Kurdish protests in Iraq
2011 Mangystau riots
2011 Northern Ireland riots
2011 Oakland general strike
2011 Qianxi riot
2011 Shanghai riot
2011 Vancouver Stanley Cup riot
2011 Zengcheng riot
2011 Zhili riot
2011 Zhongshan riot
2011 attack on the British Embassy in Iran
2011 attack on the Israeli Embassy in Egypt
2011–2012 Bahraini uprising
2011–2012 Damascus clashes
2011–2012 Daraa Province Clashes
2011–2012 Idlib Governorate clashes
2011–2012 Libyan factional fighting
314 Taipei protest
4th Summit of the Americas
70 Million Steps Against Coups
Ahmad Al-Shahhat
Ahmed Rifaat (judge)
Aleppo protests
Alliance of Yemeni Tribes
Anti-Gaddafi forces
April 2009 Thai political unrest
April 6 Youth Movement
Ash-shab yurid isqat an-nizam
Assassination of Benazir Bhutto
BART Police shooting of Oscar Grant
Bahrain: Shouting in the Dark
Battle of Douma
Battle of Sana'a (2011)
Battle of Ta'izz
Battle of Zabadani
Battle of Zinjibar
Belgrade anti-gay riot
Berkeley oak grove controversy
Best Bakery case
Black Friday (2005)
Bloody Thursday (2011)
Bronze Night
Casualties of the 2011 Libyan civil war
Casualties of the 2011–2012 Bahraini uprising
Coalition of Socialist Forces
Coalition of the Youth of the Revolution
Constituent Assembly of Tunisia
Copenhagen December Riot
Death of Ahmed Jaber al-Qattan
Death of Ali Abdulhadi Mushaima
Death of Ali Jawad al-Sheikh
Death of Fadhel Al-Matrook
Death of Hamza Ali Al-Khateeb
Death of Ian Tomlinson
Death of Khaled Mohamed Saeed
Death of Muammar Gaddafi
December 2011 Syrian–Turkish border clash
Dissenters March
Egyptian Tank Man
El-Ghad Party
February 2010 Australian cyberattacks
February 28, 2004 hand-in-hand rally
Free Libyana
Free Syrian Army
Great American Boycott
Guantánamo Bay hunger strikes
Gulbarg Society massacre
Homs airbase ambush
Homs massacre
Houphouët-Boigny Arena stampede
Human Price of Freedom and Justice
Human rights in Egypt under the Supreme Council of the Armed Forces
Human rights reports on 2011–2012 Bahraini uprising
International reactions to the 2011 Egyptian revolution
International reactions to the 2011 Libyan civil war
International reactions to the 2011 military intervention in Libya
International reactions to the 2011–2012 Bahraini uprising
International reactions to the 2011–2012 Syrian uprising
International reactions to the death of Muammar Gaddafi
January 19th incident
January 2012 al-Midan bombing
July 2009 French riots
July 2009 Ürümqi riots
Kosovo is Serbia protest 2008
Liberty Street Protest
Libya Contact Group
Libyan civil war
Local Coordination Committees of Syria
London Conference on Libya
Macassar Village Land Occupation
Manama Paper
Maspero demonstrations
Mihalis Filopoulos
Mohamed Bouazizi
Montebello High School flag flipping incident
N2 Gateway occupations
Navi Mumbai Holi riots
New Castle Correctional Facility
Occupy Ashland
Occupy Atlanta
Occupy Austin
Occupy Baltimore
Occupy Berlin
Occupy Boston
Occupy Buffalo
Occupy Canada
Occupy Charlotte
Occupy Chicago
Occupy Cincinnati
Occupy Cork
Occupy Dallas
Occupy Dame Street
Occupy Edinburgh
Occupy Eugene
Occupy Glasgow
Occupy Houston
Occupy Las Vegas
Occupy London
Occupy Melbourne
Occupy Nashville
Occupy Oakland
Occupy Philadelphia
Occupy Pittsburgh
Occupy Portland
Occupy Providence
Occupy Redwood City
Occupy Sacramento
Occupy Salem
Occupy San Diego
Occupy San Francisco
Occupy San Jose
Occupy Seattle
Occupy St. Louis
Occupy Sydney
Occupy Texas State
Occupy Toronto
Occupy Wall Street
Occupy Windsor
October 2011 Jabal al-Zawiya clashes
Operation Egypt
Poor People's World Cup
Port Said Stadium clashes
Protest in South Africa

Protests of Wukan
Refugees of the 2011–2012 Syrian uprising
Rif Dimashq blockades
River Run Riot
San Bernardino punk riot
San Luis Obispo Mardi Gras controversy
Saudi Arabian municipal elections, 2011
Science Is Vital rally
Shabeeha
Shaoguan incident
Shishou incident
Siege of Baniyas
Siege of Daraa
Siege of Deir ez-Zor
Siege of Hama
Siege of Homs
Siege of Jisr al-Shughur
Siege of Latakia
Siege of Rastan and Talbiseh
Siege of Talkalakh
Social situation in the French suburbs
Socialist Popular Alliance Party
Society for Development and Change
Strategy-31
Symphony Way Pavement Dwellers
Syrian National Council
SyrianFreePress
Tea Party movement
Tea Party of Nevada
Timeline of 2011 England riots
Timeline of the 2009 Iranian election protests
Timeline of the 2011 Egyptian revolution
Timeline of the 2011 Egyptian revolution under Hosni Mubarak's rule
Timeline of the 2011 Egyptian revolution under Supreme Council of the Armed Forces
Timeline of the 2011 Libyan civil war and military intervention (16 August – 23 October)
Timeline of the 2011 Libyan civil war and military intervention (19 March – May)
Timeline of the 2011 Libyan civil war and military intervention (June – 15 August)
Timeline of the 2011 Libyan civil war before military intervention
Timeline of the 2011 military intervention in Libya
Timeline of the 2011–2012 Bahraini uprising
Timeline of the 2011–2012 Saudi Arabian protests
Timeline of the 2011–2012 Saudi Arabian protests (January–April 2011)
Timeline of the 2011–2012 Saudi Arabian protests (May–December 2011)
Timeline of the 2011–2012 Saudi Arabian protests (from January 2012)
Timeline of the 2011–2012 Syrian uprising (May–August 2011)
Timeline of the 2011–2012 Syrian uprising (September–December 2011)
Timeline of the 2011–2012 Syrian uprising (from January 2012)
Timeline of the 2011–2012 Yemeni uprising (23 September – December 2011)
Timeline of the 2011–2012 Yemeni uprising (3 June – 22 September 2011)
Timeline of the 2011–2012 Yemeni uprising (January – 2 June 2011)
Timeline of the 2011–2012 Yemeni uprising (from January 2012)
Torture during the 2011–2012 Bahraini uprising
Trial of Lex Wotton
Trials and judicial hearings following the 2011 Egyptian revolution
Tunisian revolution
Turn Your Back on Bush
USAbilAraby
Umma Islamic Party
United Nations Security Council Resolution 1973
Venezuelan general strike of 2002–2003
Vita Cortex sit-in
Whites Only Scholarship
Workers Democratic Party
Wrocław football riot 2003
Xenophobia in South Africa
Zakariya Rashid Hassan al-Ashiri

BIBLIOGRAPHY

[1] The Economist. *The people formerly known as the audience*. URL: http://www.economist.com/node/18904124 (visited on 08/10/2011).

[2] The Economist. *Protest in Egypt: Another Arab regime under threat*. 2011. URL: http://www.economist.com/node/18013760.

[3] sueddeutsche.de. *Krise in Ägypten - Die Kinder des 6. April und der Tag der Entscheidung - Politik*. 2011. URL: http://www.sueddeutsche.de/politik/krise-in-aegypten-die-kinder-des-april-rufen-zum-protest-1.1053426.

[4] Wikipedia. *Revision history of 2011 Egyptian revolution*. URL: http://en.wikipedia.org/w/index.php?title=2011_Egyptian_revolution&dir=prev&action=history.

[5] J. Giles. „Internet encyclopaedias go head to head." In: *Nature* 438.7070 (2005), pp. 900–901. ISSN: 0028-0836.

[6] Wikipedia. *Wikipedia Page Views*. URL: http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm (visited on 08/10/2011).

[7] A. Chadwick. *Routledge handbook of Internet politics*. Taylor & Francis, 2009. ISBN: 0203962540.

[8] The Economist. *Libya: A civil war beckons*. 2011. URL: http://www.economist.com/node/18290470.

[9] B. Suh et al. „Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations." In: *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*. IEEE. 2007, pp. 163–170.

[10] F.Å. Nielsen. „Wikipedia research and tools: Review and comments." In: (2011).

[11] Wikipedia. *Statistics*. URL: http://stats.wikimedia.org/EN/Sitemap.htm (visited on 08/10/2011).

[12] Wikipedia. *History of Wikipedia*. URL: http://en.wikipedia.org/wiki/History_of_Wikipedia (visited on 12/10/2011).

[13] *Wikimedia Foundation*. URL: http://en.wikipedia.org/wiki/Wikimedia_Foundation (visited on 01/24/2012).

[14] *Wikimedia Foundation annual report 2010-2011*. URL: https://upload.wikimedia.org/wikipedia/commons/4/48/WMF_AR11_SHIP_spreads_15dec11_72dpi.pdf (visited on 01/24/2012).

[15] Wikipedia. *Protection policy*. URL: http://en.wikipedia.org/wiki/Wikipedia:Protection_policy (visited on 11/16/2011).

[16] *Minor edit*. URL: http://en.wikipedia.org/wiki/Help:Minor_edit (visited on 01/27/2012).

[17] Wikipedia. *Why create an account?* URL: http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account%3F (visited on 08/10/2011).

[18] Fabian Kaelin. *Research:Anonymous edits*. URL: http://meta.wikimedia.org/wiki/Research:Anonymous_edits (visited on 12/10/2011).

[19] *Wikipedia Bots*. URL: http://en.wikipedia.org/wiki/Wikipedia:Bots (visited on 12/10/2011).

[20] F.B. Viégas et al. „Talk Before You Type: Coordination in Wikipedia." In: *Proceedings of HICSS*. Vol. 40. 2007.

[21] M. Kramer, A. Gregorowicz, and B. Iyer. „Wiki trust metrics based on phrasal analysis." In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–10.

[22] B.T. Adler and L. De Alfaro. „A content-driven reputation system for the Wikipedia." In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 261–270.

[23] B.T. Adler et al. „Assigning trust to wikipedia content." In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–12.

[24] D. Hardy. „Volunteered geographic information in Wikipedia." PhD thesis. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011.

[25] J.A. Muir and P.C.V. Oorschot. „Internet geolocation: Evasion and counterevasion." In: *ACM Computing Surveys (CSUR)* 42.1 (2009), p. 4.

[26] M.D. Lieberman and J. Lin. „You are where you edit: Locating Wikipedia users through edit histories." In: *ICWSM'09* (2009), pp. 106–113.

[27] D.K. Lewis. *Philosophical papers*. Vol. 2. Oxford University Press, USA, 1987.

[28] E. Adar, M. Skinner, and D.S. Weld. „Information arbitrage across multi-lingual Wikipedia.“ In: *Proceedings of the second ACM international conference on Web search and data mining*. ACM. 2009, pp. 94–103.

[29] B. Hecht and D. Gergle. „The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context.“ In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 291–300.

[30] A. Kittur, E.H. Chi, and B. Suh. „What's in Wikipedia?: mapping topics and conflict using socially annotated category structure.“ In: *Proceedings of the 27th international conference on Human factors in computing systems*. ACM. 2009, pp. 1509–1512.

[31] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. „Studying cooperation and conflict between authors with history flow visualizations.“ In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 575–582. ISBN: 1-58113-702-8. DOI: http://doi.acm.org/10.1145/985692.985765.

[32] A. Kittur et al. „Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie.“ In: *World Wide Web* 1.2 (2007), p. 19.

[33] A. Kittur et al. „He says, she says: Conflict and coordination in Wikipedia.“ In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 453–462.

[34] S.K. Card, J.D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[35] *List of Wikipedias*. URL: http://en.wikipedia.org/wiki/List_of_Wikipedias (visited on 01/30/2012).

DECLARATION

I declare that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other University.

Formulations and ideas taken from other sources are cited as such.

Berlin, 2012–02–04

David Kaltschmidt

B

ERRATA

The following mistakes need to be corrected from the version that I submitted as my diploma thesis:

B.1 MINOR

- In APPARATUS (ch. 4), the 4th paragraph should read: "The application can be accessed over the internet."
- In Proximity metrics (subsec. 5.2.1), in the last paragraph, the mean share of the located text is actually 26.2%, instead of 23.8%.
- In Hypotheses analysis (sec. 5.5), the captions for figures 22, 23, and 24 should all contain "(1.0 = 100)", without the percent sign, since the values represent index numbers.
- In RESULTS (ch. 6), the first sentence should read: "This chapter presents the quantitative analysis of a set of articles of political events from the English Wikipedia."
- In Data set preparation (subsec. 6.1.2), the first sentence of the second paragraph should read: "Then, all articles in CONFLICTSRAW are tested whether they fulfill the Article requirements for an analysis."
- In Characteristics of the set (sec. 6.2), in table 3, the footnote [a] is missing and should read: "Excluding bots."
- In H5 – H8: Participation (sec. 6.4) and H9 – H10: Text survival (sec. 6.5), all mentions of *mean* or *average* should read "median" as they refer to the box plots' middle lines, see figures 32 and 33.

B.2 OMISSIONS

In Implementation of the signature distance (subsec. 4.4.6.3), listing 9 presents an algorithm to efficiently calculate the signature distance for all revisions of an article. It uses an optimization called memoization for which a recursive formula is being used. The step of converting the iterative declaration of the signature

distance $D(\alpha)$ into a recursive declaration was omitted and is added here for completion.

Using the same notation, let α be an article, ρ be an author, and π be a revision. For a sample of articles S let A be the set all articles in the sample, P be the set of all located authors in the sample, and Π the set of all revisions in the sample.

$$A = \{\alpha : \alpha \in S\} \qquad P = \{\rho : \rho \in S\} \qquad \Pi = \{\pi : \pi \in S\}$$

Further, let $\eta(\rho, \alpha)$ denote the contribution(s) of author ρ to article α. Then, $N(\alpha)$ are all contribution(s) to article α while $P(\alpha)$ are the author(s) who have made contributions to article α:

$$N(\alpha) = \{\eta(\rho, \alpha) : \rho \in P\} \qquad P(\alpha) = \{\rho : \rho \in N(\alpha)\}$$

In addition, let $\gamma(\pi, \alpha)$ denote whether π is a revision of article α. Then, $\Gamma(\alpha)$ are all revisions to article α:

$$\Gamma(\alpha) = \{\gamma(\pi, \alpha) : \pi \in \Pi\}$$

Since a contribution creates a revision, the total number n of contributions equals the number of revisions to article α:

$$n = |\Gamma(\alpha)| = |N(\alpha)|$$

Let further $\delta(\rho, \alpha)$ be the geodesic distance between author ρ and article α. Additionally, let $\zeta(\pi)$ denote the author ρ of revision π to article α. It follows that the distance of a revision to the article is the same as the distance between the article and that revision's author:

$$\rho = \zeta(\pi) \;\Rightarrow\; \delta(\zeta(\pi), \alpha) = \delta(\rho, \alpha)$$

In the original signature distance $D(\alpha)$, all contributions by an author were summed up by $\eta(\rho, \alpha)$. Considering each of these contributions as a single revision π, the signature distance can also be expressed as follows:

$$\begin{aligned} D(\alpha) &= \sum_{\forall \rho \in P(\alpha)} \frac{|\eta(\rho, \alpha)| \cdot \delta(\rho, \alpha)}{|N(\alpha)|} \\ &= \sum_{\forall \pi \in \Gamma(\alpha)} \frac{\delta(\zeta(\pi), \alpha)}{|\Gamma(\alpha)|} \end{aligned}$$

When expanded, it becomes obvious that the signature distance for $n = |\Gamma(\alpha)|$ revisions contains the signature distance for the previous $n - 1$ revisions:

$$
\begin{aligned}
D(\alpha) &= \frac{\delta(\zeta(\pi_1),\alpha) + \ldots + \delta(\zeta(\pi_{n-1}),\alpha) + \delta(\zeta(\pi_n),\alpha)}{n} \\
&= \frac{\delta(\zeta(\pi_1),\alpha) + \ldots + \delta(\zeta(\pi_{n-1}),\alpha)}{n} + \frac{\delta(\zeta(\pi_n),\alpha)}{n} \\
&= \frac{(\delta(\zeta(\pi_1),\alpha) + \ldots + \delta(\zeta(\pi_{n-1}),\alpha))(n-1)}{n(n-1)} + \frac{\delta(\zeta(\pi_n),\alpha)}{n} \\
&= \frac{n-1}{n} \cdot \frac{\delta(\zeta(\pi_1),\alpha) + \ldots + \delta(\zeta(\pi_{n-1}),\alpha)}{n-1} + \frac{\delta(\zeta(\pi_n),\alpha)}{n}
\end{aligned}
$$

Hence, the signature distance can be restated recursively:

$$D(\alpha,\pi_1) = \delta(\zeta(\pi_1),\alpha) \tag{B.1}$$

$$D(\alpha,\pi_n) = D(\alpha,\pi_{n-1}) \cdot \frac{n-1}{n} + \frac{\delta(\zeta(\pi_n),\alpha)}{n} \tag{B.2}$$

The distance for the first revision, B.1, is the base case for which the signature distance of the article is the distance between the author ρ of revision π_1 and the article α. For subsequent revisions, the signature distance is calculated using the result for the previous revision, B.2.

B.3 QUANTITATIVE ANALYSIS

A postmortem of the quantitative analysis data, used in RESULTS (ch. 6), revealed that around 15% of articles were illegitimately rejected[1].

A re-run of the analysis was done on 02/07/2012[2]. There, a net[3] five articles were added to the categories and templates that made up the set CONFLICTSRAW:

1981 General strike in Bielsko-Biała	February 2012 bombardment of Homs
2011-2012 Syrian civil war	Great Nordic Biker War
Castellammarese War	Masrena

The new raw set CONFICTSRAW-B consists of 746 articles. I denote these new articles together with the previously rejected articles as SKIPPED. The resulting set for analysis CONFLICTS-B thus contains the articles of CONFLICTS plus the articles of SKIPPED. The articles in SKIPPED are:

1 These articles passed most of the initial Article requirements (subsec. 4.4.1). But then, due to a Toolserver unavailability, the authorship could not be determined and the article was rejected.

2 The final run for the thesis was done on 02/05/2012.

3 Two old articles, *Homs Bombardment (2012)* and *Homs massacre* were merged into *February 2012 bombardment of Homs*.

2006 Dalit protests in Maharashtra
2006 Indian anti-reservation protests
2006 Islamist demonstration outside the Embassy of Denmark in London
2006 Oaxaca protests
2006 protests in Hungary
2006 student protests in Chile
2006 youth protests in France
2007 Bersih rally
2007 Burmese anti-government protests
2007 HINDRAF rally
2007 MacArthur Park rallies
2007 Macau labour protest
2007 Macau transfer of sovereignty anniversary protest
2007 Russian protests
2008 Armenian presidential election protests
2008 Cameroonian anti-government protests
2008 Icelandic lorry driver protests
2008 Otago NORML protests
2008–2010 Thai political crisis
2009 Ashura protests
2009 G-20 London summit protests
2009 Georgian demonstrations
2009 Guinea protest
2009 Luquan protest
2009 May Day protests
2009 Tamil diaspora protests in Canada
2009 student protests in Austria
2009 student protests in Croatia
2009–2010 Iranian election protests
2010 Air Untuk Rakyat rally
2010 Canada anti-prorogation protests
2010 Catalan autonomy protest
2010 G-20 Toronto summit protests
2010 Macau labour protest
2010 Macau transfer of sovereignty anniversary protest
2010 Thai political protests
2010 Xinfa aluminum plant protest
2010–2011 Greek protests
2010–2011 Senegalese protests
2010–2012 Algerian protests
2011 Albanian opposition demonstrations
2011 Armenian protests
2011 Chilean protests
2011 Chinese pro-democracy protests
2011 Djiboutian protests
2011 Indian anti-corruption movement
2011 Iranian protests
2011 Iraqi protests
2011 Israeli social justice protests
2011 Kurdish protests in Turkey
2011 Lebanese protests
2011 London anti-cuts protest
2011 Malawian protests
2011 Maldivian protests
2011 Mexican protests
2011 Nanchang mass suicide protest
2011 Omani protests
2011 Rome demonstration
2011 Spanish protests
2011 Sudanese protests
2011 United Kingdom anti-austerity protests
2011 United States public employee protests
2011 Western Saharan protests
2011 Wisconsin protests
2011 Yunnan protest
2011–2012 Jordanian protests
2011–2012 Kuwaiti protests
2011–2012 Moroccan protests
2011–2012 Saudi Arabian protests
2011–2012 Yemeni uprising
2012 Romanian protests
517 Protest
Aboriginal Day of Action
Anti-Roma protests in Bulgaria 2011
Arab Spring
Berkeley Marine Corps Recruiting Center protests
Cottage cheese boycott
Cow head protests
Crimean anti-NATO protests of 2006
Dazhou protests of 2007
December 2005 protest for democracy in Hong Kong
Dongzhou protests
February 2012 bombardment of Homs
Glasgow school closures protest, 2009
January 27, 2007 anti-war protest
March 19, 2008 anti-war protest
May 2007 RCTV protests
Mid 2011 Telangana protests
November 15, 2008 anti-Proposition 8 protests
Occupy Charlottesville
Occupy Nigeria
Occupy movement
Port of Tacoma protests, March 2007
Protests following the 2011 Russian elections
Protests regarding 2008 South Ossetia war
Republic Protests
September 15, 2007 anti-war protest
Sharad Pawar slapping incident
SlutWalk
Taxpayer March on Washington
Tea Party protests
Tripoli protests and clashes (February 2011)
Zakariya Rashid Hassan al-Ashiri
Zhejiang solar panel plant protest

The new set CONFLICTS-B contains 380 articles. Although bigger in size, the new set's effect on the outcome of the quantitative analysis is negligible. This is due to the fact that articles are shuffled within the set after retrieval.[4] The same metrics as in table 3 have been measured for CONFLICTS-B. For a comparison, the mean values for both sets are shown in table 4.

At first glance, the articles in SKIPPED seem to contribute especially old articles to CONFLICTS-B. These older articles tend to have a higher number of revisions as well as contributors, raising the mean values for age, revisions, and contributors.

Regarding the hypothesis tests, the results for CONFLICTS-B differ only for H3, H9, and H10. Of the 106 articles that qualified for H3, 53% (previously 64%) were created first in the English Wikipedia. A number of languages editions for first article creation were added by SKIPPED. For non-native English authors, Arabic remained the second most popular language edition to publish an article in, see figure 35.

In table 4, the means of the proximity metrics for all last revisions of the set show a slight decrease. Therefore, a bigger proportion of content in the articles in SKIPPED is local when

4 As a result, the run on CONFLICTS was done, in effect, on a random sample.

Table 4: Metrics (mean) for articles in CONFLICTS and CONFLICTS-B

	CONFLICTS	CONFLICTS-B
Age	693.6d	728.2d
Size (kB)	26.3	28.9
Revisions	303.5	357.7
Contributors[a]	81.8	95.9
– anonymous	37.3	44.2
Located contributors (%)	49.8	48.7
– anonymous (%)	40.0	41.5
– by PMCU (%)	5.2	4.2
– by user page (%)	4.6	3.0
Cumulative		
Unique countries of origin	8.8	9.6
Signature distance (km)	4,271.3	4,252.7
Present in latest revision		
Unique countries of origin	5.4	5.9
Signature dist. e.surv (km)	4,465.1	4,330.4
Signature dist. t.surv (km)	4374.6	4,256.4
e.surv index	105.6	101.7
t.surv index	106.3	103.8
Located text (%)	33.9	33.3

[a] Excluding bots.

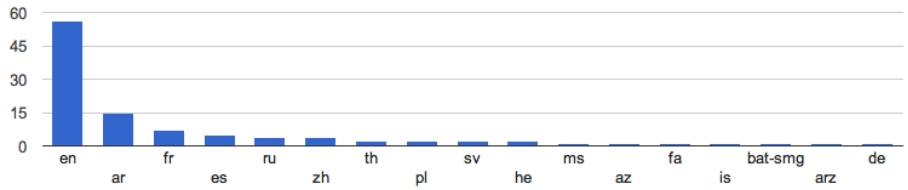
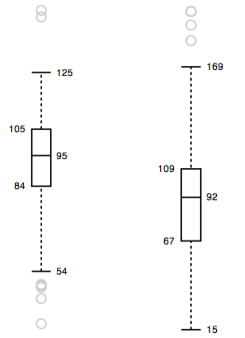


Figure 35: Language editions chosen for first article in CONFLICTS-B



(a) $\overline{I_{e.surv}}$ (b) $\overline{I_{t.surv}}$

Figure 36: Distribution of means of localness indexes in CONFLICTS-B

compared to CONFLICTS. Figure 36 shows the distribution of the means of the localness indexes for all articles. The distribution of means displays slightly lower values than the same distributions for CONFLICTS, see figure 34.