Title this and that

Process Book

Alex Mocanu, Manana Lortkipanidze

Data Visualization, EPFL 2018

Contents

[Introduction 3](#_Toc533000758)

[Overview 3](#_Toc533000759)

[Motivation 4](#_Toc533000760)

[Target Audience 4](#_Toc533000761)

[Inspiration 4](#_Toc533000762)

[Dataset 5](#_Toc533000763)

[Exploratory Data Analysis 5](#_Toc533000764)

[Popularity of Articles 5](#_Toc533000765)

[Spikes in Visitor Counts 5](#_Toc533000766)

[Concept 6](#_Toc533000767)

[Designs 6](#_Toc533000768)

[Initial Idea 6](#_Toc533000769)

# Introduction

This process book serves as a description of our project for data visualization (COM-480) class at EPFL, Fall 2018. The aim of the class was to introduce us to various informative/powerful/state of the art data visualization techniques and methods, while emphasizing not just on technical details or information they convey, but also on esthetic, artsy side of visualization. Data Analysis heavily relies on data visualization and material learnt is crucial for anyone interested in the field.

This process book explains reasoning behind every analysis or design step of our project starting from the initial abstract idea, drafts and sketches all the way through the final product. This process book guides its readers through our thought process from the day first, providing along insights into dataset used, design choices made, explaining functionality of visualization, providing details of its implementation and presenting usage scenarios. Nevertheless, discussing possible impacts and future improvements as well as concrete use-case scenarios of our visualizations. Finally, Technical details will be provided for people interested in further development/re-usage of our work.

## Overview

Wikipedia, supported and owned by Wikimedia foundation (non-profit organization), is not only the most widely used online encyclopedia, it is also considered as one of the most visited/popular websites in the world. Wikimedia foundation operates through donations received by people willing to support them, meaning that Wikipedia is free of charge for everyone.

January 15th of 2001 was the day Wikipedia was launched for the first time by Jimmy Wales and Larry Sanger. It was and is based on a model of openly editable and viewable content. The website contains around 5,769,478 articles in total and is multilingual.

There exists a controversy regarding accuracy of Wikipedia articles, since some accused it for exhibiting bias and presenting a mixture of "truths, half-truths, and some falsehoods". However, review published in 2005 concluded that accuracy of Wikipedia articles is almost equal to the accuracy of Britannica. Therefore, Wikipedia can be considered as the biggest and possibly the best encyclopedia in the world.

## Motivation

Our initial motivation was to study human behaviour based on Wikipedia data set. As already mentioned above, Wikipedia is a free and openly editable Encyclopaedia. It is created by volunteers and is one of the most popular websites. Therefore, political news, climatic changes, new trends, technologies, movies and basically all-important life events impact its content, edit requests, visitor counts and frequencies. Therefore, we thought it would provide good insight into human mindset and behaviour patterns.

Question we asked were the following: how can we define, more specifically quantify popularity of articles? How can we measure spikes in visitor counts? Can we identify major events based just on sudden peaks in number of visitors for a specific page? Which pages get affected for specific category of event? How does increase in visitors on pages affect visitor count of its hyperlinked pages? More specifically, how far does the interest spread in terms of related pages? Finally, and most importantly (scope of our course project), how can we best visualize the answers to the questions provided above?

Emphasized will be visualizations depicting popularity and its spread along network, as well as spike linkage with (major) life events and effect on its neighbours. These two visualizations are result of our project that we will discuss in fine detail in the following sections.

## Target Audience

Our target audience is anyone interested in human mindset and reflection of real-life events and trends on human behaviour regarding Wikipedia. Our target audience are people, curious to see how powerful/widely used Wikipedia articles are and consequently how one can discover patterns of everyday life events through Wikipedia visitor counts or alternatively observe impact of specific events captured by the number of visitors, hence people who were affected, reached or interested by a specific event/news.

## Inspiration

[ToDo]

# 

# Dataset

We used Wikipedia dataset originally retrieved from SNAP repository. This is a website containing human navigation paths on Wikipedia that were collected through human-computer interaction game called Wikispeedia. In the game, players received pairs of Wikipedia articles and had to reach one from the other through Wikipedia links between them. The dataset provided consists of 4604 different articles. For the purpose of our project we were provided with cleaned and extracted dataset from the above-mentioned website, consisting of names of articles, their categories and linkage information in the form of an adjacency matrix.

Since the goal of our project was to identify popularity, spikes in popularity and link those occurrences with the corresponding events in real world for visualization, we could have used an arbitrary sample from the millions of Wikipedia articles as long as we could identify unusual activities caused by external factors. Therefore, we deemed obtained articles that we were kindly provided by Mr. [???ToDo] as sufficient.

Additionally, we queried number of visits for all days in 2017 for all articles in our dataset from Wikipedia API. This information served as basis for popularity and spike detection

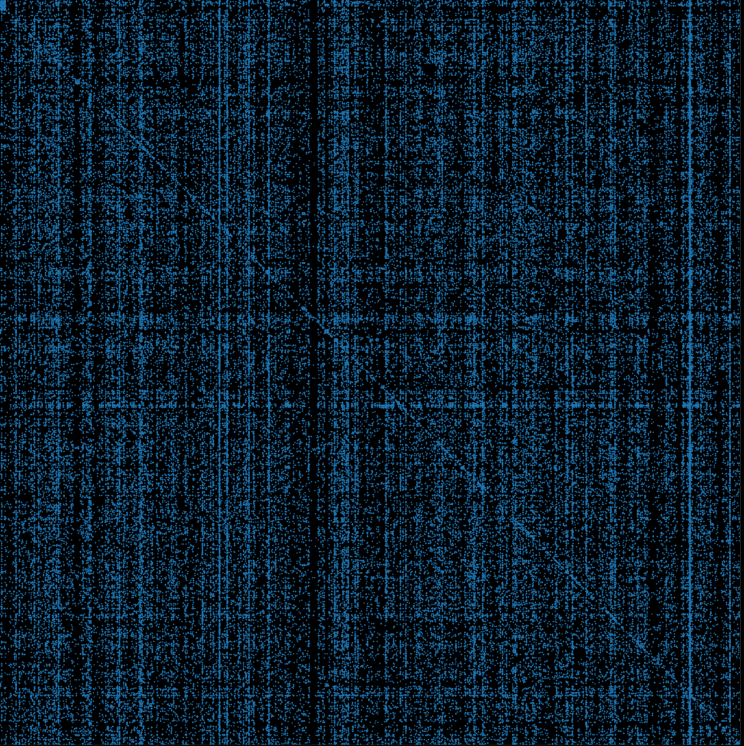
Prior to discussing design choices, we think we should briefly introduce main characteristics and descriptive statistics of our data for the purpose of making further sections more comprehensible and insightful. As well as explain how we measure popularity and identify articles with unusually high visitor counts compared to their baseline.

## Exploratory Data Analysis

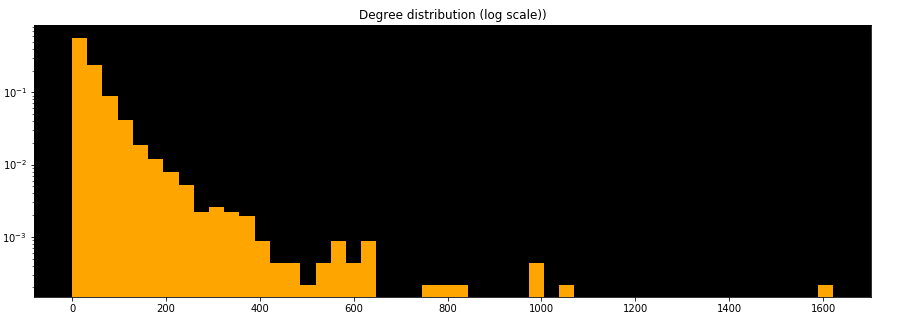
As mentioned above, SNAP repository dataset contains 4604 articles. However, some of them (less than 0.5%) were removed because they were duplicated when representing different categories. As for the categories, they are hierarchical, and we have 15 top level categories in total. It must be noted, that one article can belong to several categories, including several top-level categories. Distribution of articles over categories is as follows:

[categories histogram ToDo]

We should also mention, that as expected adjacency matrix is sparse:



The adjacency matrix above forms a graph with one big connected component and 3 additional isolated articles/nodes (that were then removed from our dataset). Furthermore, observing degree distribution below, we conclude that our network resembles scale free network:



Scale free network means, that we have hubs (articles with very high degree compared to other articles) and the rest of them are sparse. Additionally, we observed that graph has small world property, since diameter of our network is 8. [Meaning information/influence spreads over all network rapidly. Therefore, it would make sense for popularity or spike to be spread to neighboring nodes. ToDo]

## 

## Popularity of Articles

Popularity of articles is determined using daily visit counts. We plan to visualize n most popular articles/nodes for a selected time frame, providing relevant statistics alongside and possibility to explore its neighboring nodes. We won’t go into details of popularity in this section, since one big part of our visualization functionality description will be devoted to the topic.

## Spikes in Visitor Counts

In order to identify spikes in visitor counts we employ different strategies. Firstly, we estimate average visit count for all articles over a year. And then identify days with unusually high average visitor counts, under the assumption that this high average number of visitors was caused by significant spikes in one or several articles. After selection days throughout the year 2017, we try to identify articles responsible for the unusually high daily average number. For that, we first create baseline for each article, by averaging their daily visitor counts over a year.

# Concept

## Designs

Implementation Model

Technical details

Implementation

Technical details