Hello I'm a title

Network Tour of Data Science Report, January 2019

Isabela Constantin, Adélie Garin, Celia Hacker, Michael Spieler

I. Introduction

Using the Wikipedia data that is now open source, we would like to understand how the number of views on some pages depend on a type of event. We will consider the two following types of events for our analysis:

1) Expected events such as elections, concerts, political events
2) Unexpected events which could be for example, death of someone famous,a coup d'état, or a natural disaster

Choosing several specific events of each type, we will consider the graph built with pages concerning the event as nodes and a two pages are linked if there is a web link from one to the other. We will use both the directed graph, in which one page which is linked to the other leads to only one direction for the concerned edge, and the undirected graph, for which there exists an edge if and only if there is one page linked to the other.

The questions we ask ourselves are the following: Does the number of views on the pages propagate in the pages in different ways depending on the type of event? If yes, how do they behave?

We will consider the quotient of the number of views of the day of the event by the number of views on the day before the event.

To analyse our data and try to answer our questions, we will first start by acquiring the data and building several graphs for each type of event. Then we explore them and try to get general properties of the graphs, building a general pipeline for our analysis. We then exploit the data, trying to test our hypothesis and finally draw conclusions.

II. Data Acquisition

We constructed the graphs by selecting the Wikipedia article corresponding to each event. We then grew the graphs around them by selecting the pages the event is linking to, and the pages linked from those. To reduce the amount of data we had, we randomly subsampled by giving a higher probability of staying in the graph to the nodes that are higher in the page, meaning they are stated early so they are more likely to be important links. The resulting graphs are directed, unweighted and connected. In addition, we found the number of views per page, that we computed for the day before the event and the day of the event. For each page will consider the quotient

$$\frac{\text{number of views the day of the event}}{\text{number of views the day before}}.$$

Taking this quotient is a sort of normalization of the number of views. The higher this value is, the more the number of views on the day of the event is big compared to the number of views of the day before. We hence have a value assigned to each node of our graphs.

III. Data Exploration

The six graphs we built are the following:

TABLE I
Our graphs

	Event	***type of event***	***Wikipedia article***
Graph 1	Expected		
Graph 2	Expected		
Graph 3	Expected		
Graph 4	Unexpected		
Graph 5	Unexpected		
Graph 6	Unexpected		

ACC stands for Average Clustering Coefficient.

We start by analysing some basic properties of each graph, such as the number of nodes, the number of edges, the diameter of the graph, the average clustering coefficient with the tools available in Networkx. The results are stated in the following table.

TABLE II
Basic properties of the graphs

Graphs	**Properties**			
	Number of nodes	***Number of edges***	***Diameter***	***ACC***
Graph 1				
Graph 2				
Graph 3				
Graph 4				
Graph 5				
Graph 6				

ACC stands for Average Clustering Coefficient.

Some other properties we would like to study are the the degree distributions, the strongly connected components of the graphs and do spectral clustering on the undirected versions of our graphs.

IV. Data Exploitation

We now come to the most important part of our analysis, which is the following: we consider the number of views as a signal on the graph. We would like to see how it behaves. In order to make a sensible comparision of number of views and how these number of views evolve for each graph we use

the quotient of the number of views before and at the day of the event. The general pipeline of our analysis is :

1)
2)

V. Conclusion

Note that if we had the computational power to build a graph with all the pages involved by all the events that we considered together, it would have been much more interesting to see how the number of views evolves locally on the graph. By doing the methods presented in this report, we can only do a "zoom-in" on a specific part of this big graph and analyse it independently of the rest.

References

[1] [2] [3]

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[3] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.