A Wikipedia Tour of Death

Or How University College Boat Club Became Popular

Authors:
Isabela Constantin
Adélie Garin
Celia Hacker
Michael Spieler

Wikipedia is, according to Wikipedia itself, the fifth most visited website and the largest encyclopaedia on the World Wide Web. Being based on a model of free and openly editable content, which generates a **huge amount of traces of activity** all around the globe, it is a **source of extremely valuable data**, which can be transformed into insights about **how people use**, edit and navigate information systems.

The Idea

- Focus on a small part of the Wikipedia graph
- Understand the behaviour of people on Wikipedia
- In the context of an unexpected event (e.g. Terror attack, natural disaster, death of a famous person, …)
- Construct a graph around that page
- Study the number of page views around the time of the event

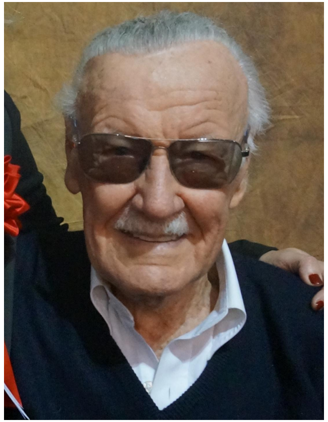
→ Death of a celebrity

Celebrities

Stephen Hawking

Physicist

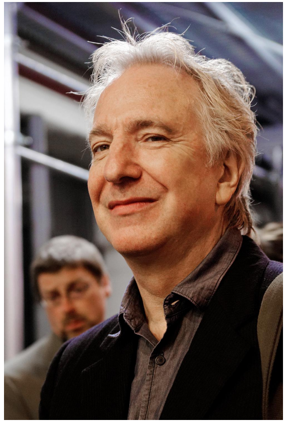
08.01.1942 - 14.03.2018



Stan Lee

Comic book writer and editor

28.12.1922 - 12.11.2018

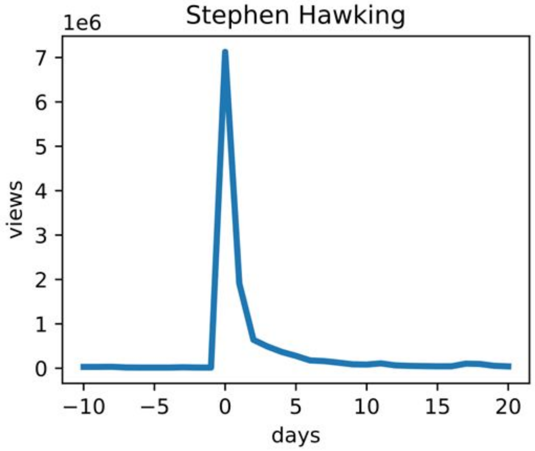


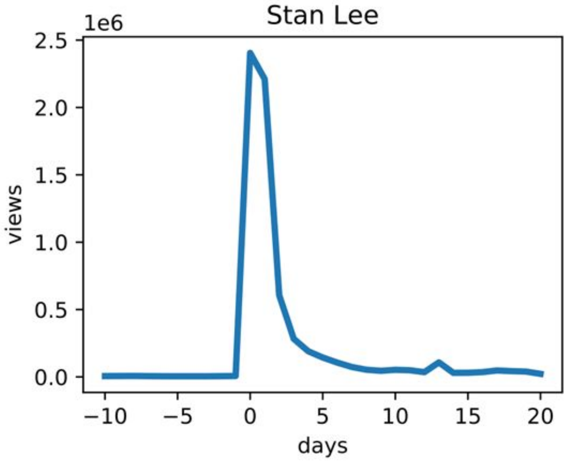
Alan Rickman

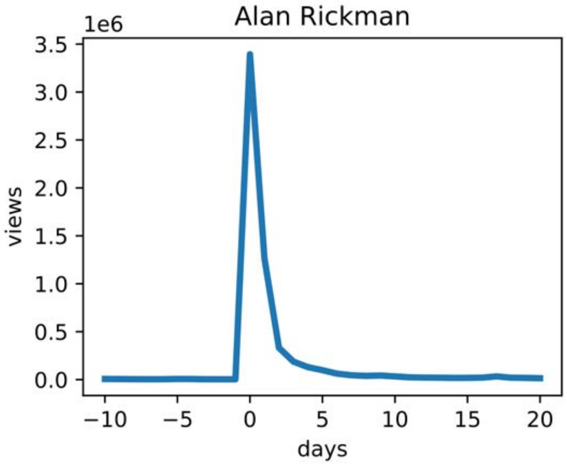
Actor

21.02.1946 - 14.01.2016

Number of views before and after they died







Data acquisition: main considerations

- Goal: **given a seed page and a date**, crawl a snapshot subgraph

- Considerations:
  - Number of nodes less than 4000, from a computation point of view
  - … but the resulting graph should be strongly related to the seed article.
  - And it should not be too sparse or too dense (varied sampling)

- Result:
  - A programmatic pipeline that with just a few function calls can crawl a graph based on given parameters.

Data acquisition in practice: Parameters

- Computation limitation:
  - breadth-first search starting from the initial node with **sampling**
  - Parameters:
    - 2701 nodes
    - Sample 150 links from if 1 hop away from the seed
    - Sample 3 links for each page if undirected link
- Meaningful graph for the seed article:
  - Non uniform sampling
  - Avoid really specific pages
- Not too sparse, not too dense:
  - Sampling from the beginning of the page induces variety in the categories of the articles.

Acquiring the node signal

- For each node, get the number of views the page had on the day before and on the day of the death
- ~~Approach 1:~~
    - Take the difference between these two numbers.
- **Approach 2:**

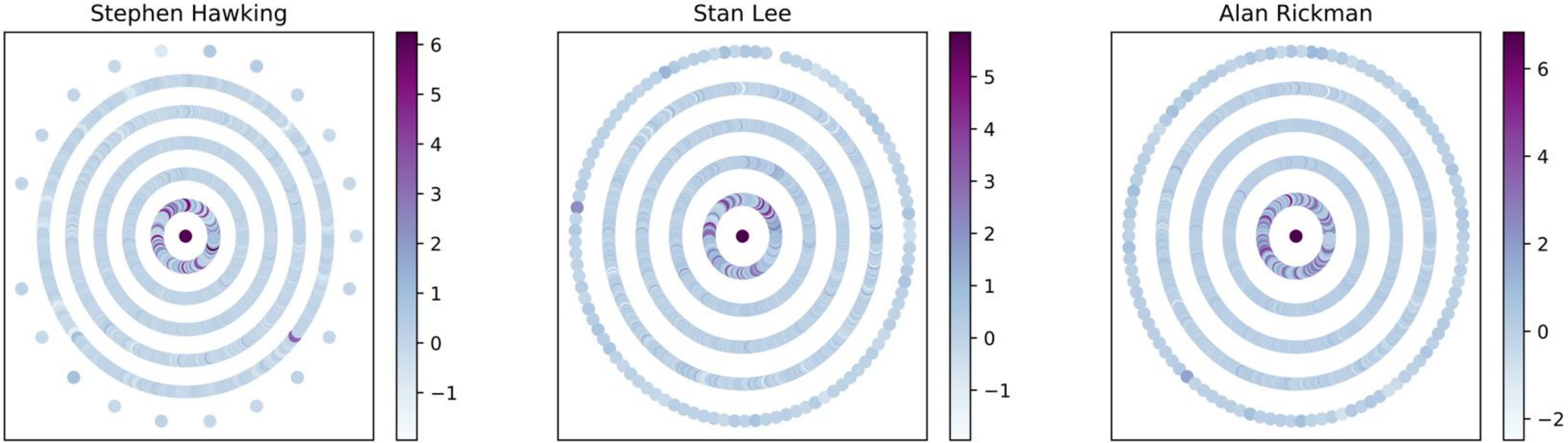
$$\text{signal} = \frac{\text{number of views on the day of the death}}{\text{number of views on the day before}}$$

- Assumption: the higher the signal, the more relevant the page to the famous person

Data Exploration: Basic Properties of the Graphs

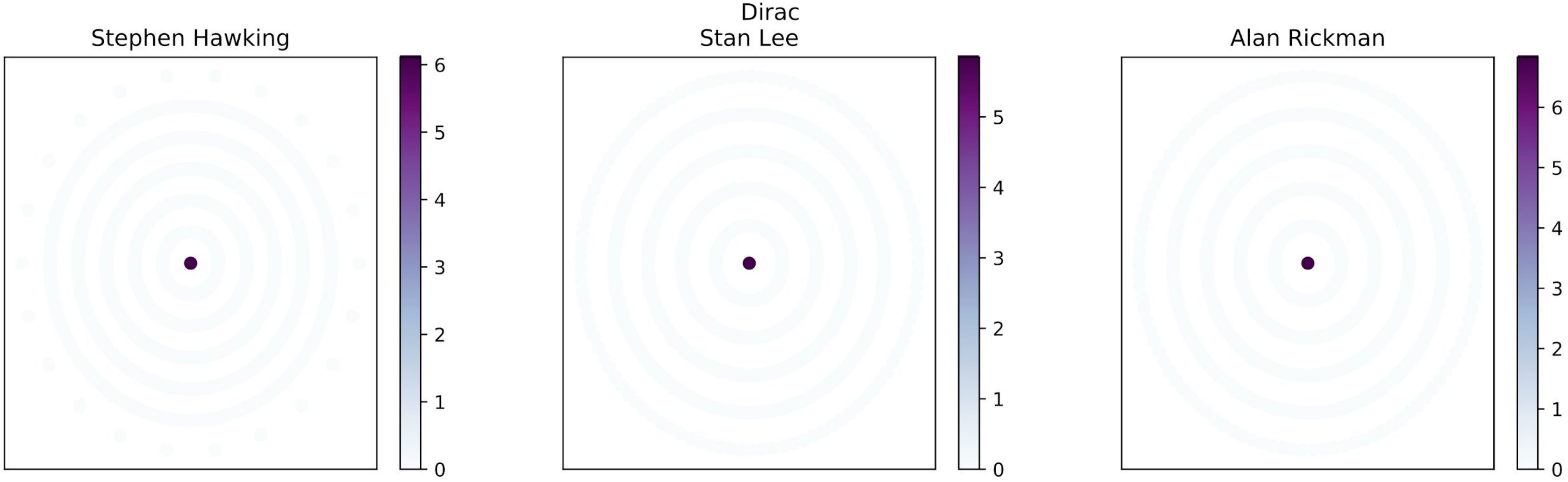
Properties	Stephen Hawking	Stan Lee	Alan Rickman
Nodes	2701	2701	2701
Edges	26921	33469	28012
Average Degree	18	21	19
Diameter	7	7	8
Average clustering coefficient	0,138	0,131	0,124
Number of triangles	41973	60096	38076
Global clustering coefficient	0,000335	0,000379	0,000251

Graph signal



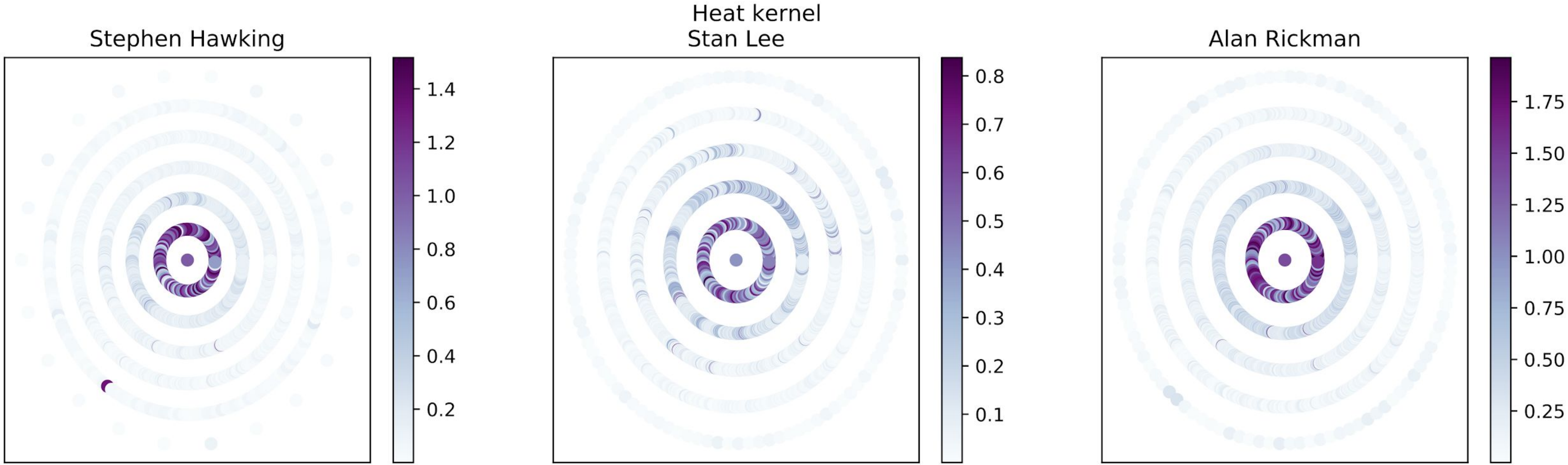
Signal model using GSP

Idea: Model the signal by **filtering** a **dirac** at the center node.



Signal model using GSP

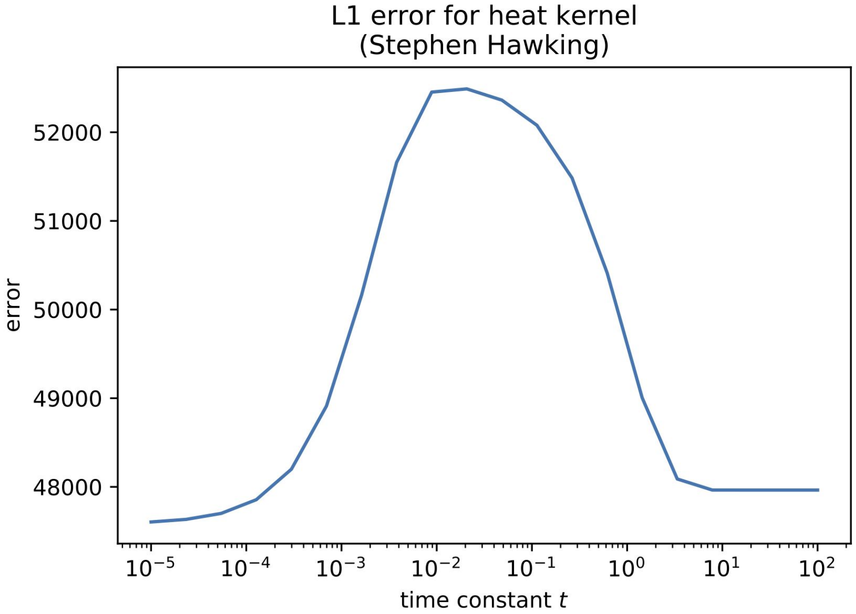
Idea: Model the signal by **filtering** a **dirac** at the center node.



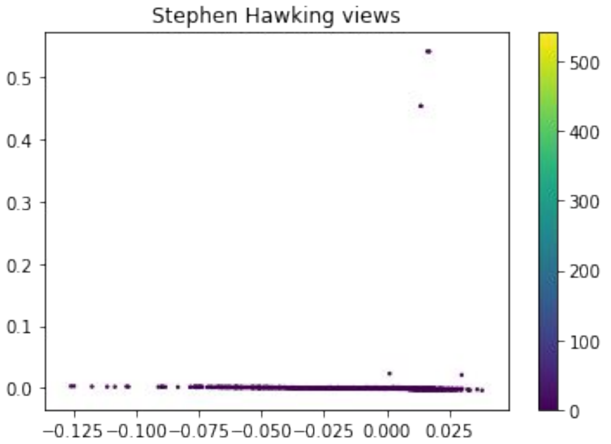
Model parameter fitting

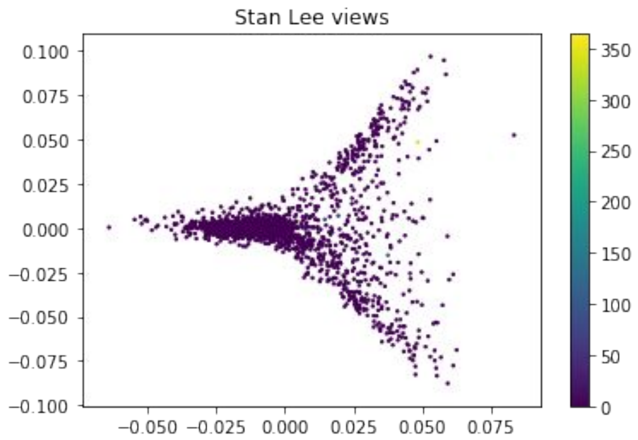
Conclusion:

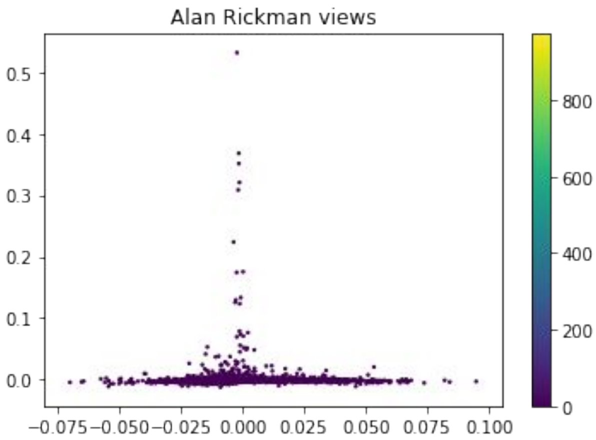
Model is not adapted to the signal.



Spectral Eigenmap

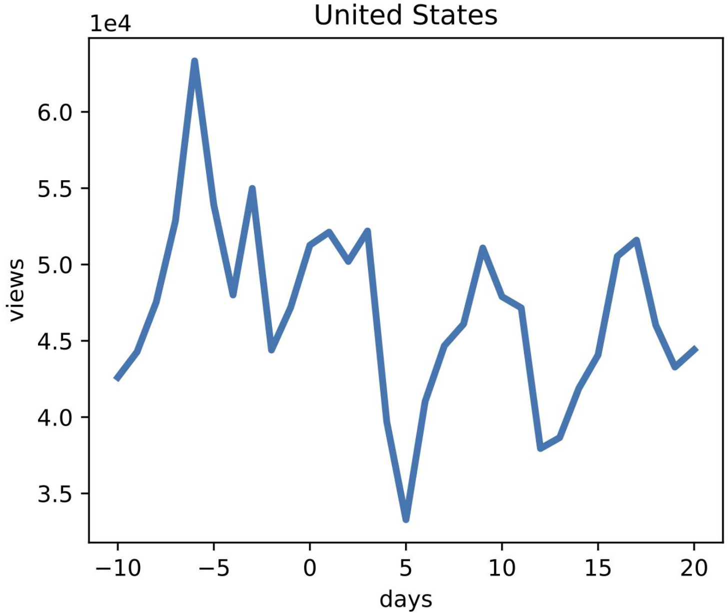






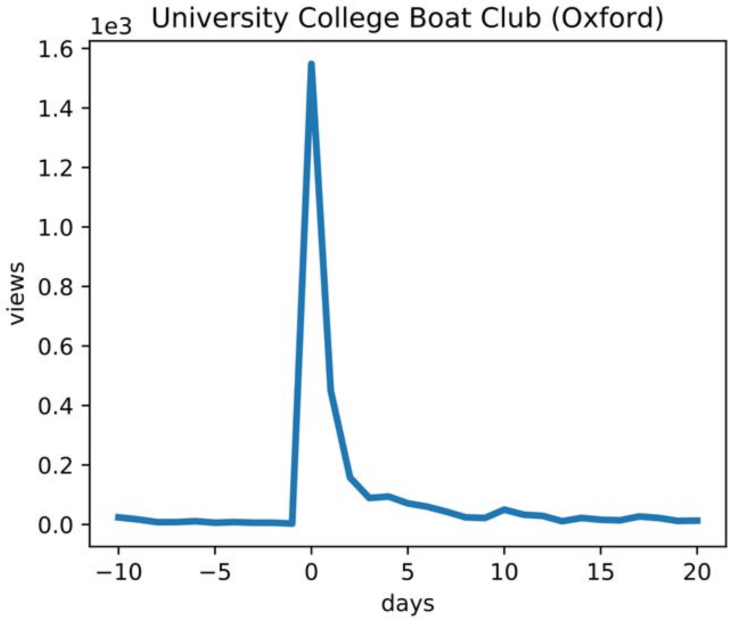
Fun Facts : outliers

- Highest degree (undirected): United States - important page in Wikipedia
- Views not affected much by Stephen Hawking, Stan Lee, Alan Rickman
- Views affected by many other factors



Fun Facts : outliers

- Highest signal: University College Boat Club: 516.00 times more views
- Views affected by death of Stephen Hawking

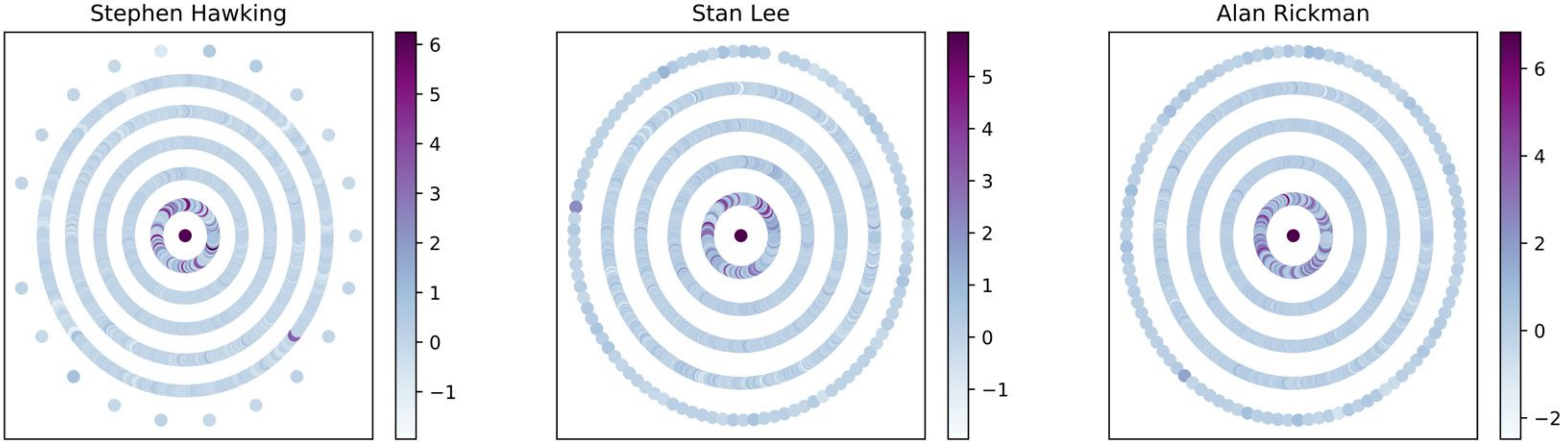


Behaviour of people on Wikipedia

Our initial assumption

- People click on a few links starting from one page
- Mostly “unexpected” pages (unknown pages VS popular pages)
- A small diffusion of the quotient signal

Conclusion



Stephen Hawking

Stan Lee

Alan Rickman

Very similar patterns between the graphs and signals
 but GSP does not seem to be the best method to simulate this behaviour

Conclusion

- Our signal does not diffuse much (beside exceptions)
- Death of famous people affect a lot “unpopular” pages
- Correlation VS Causality
- More data & computational power → better overview → better analysis

Thank you!