Human digital traces

mobility, sewage and social media

Dr. Francesco Durazzi

francesco.durazzi2@unibo.it

Looking for traces

- Information about human behavior is useful in many research field, but:
- Privacy concerns
- High costs

The *digital era* offers new ways to get a statistical glimpse on human-related data -> **it's not psicology**

Trace: a measurable quantity which describes the subject behaviour (without necessarily the intent of the subject)

Disclaimer: a health perspective

Focus on epidemiological surveillance:

Pre-warning about pathogens

Quantifying the impact on society

Predicting possible scenarios

Measure the effectiveness of control measures

How to? **Systematic collection, analysis and interpretation of data**

e.g. genomics, animal mobility flows, opinions on online social networks…

DIGITAL TRACES
Characteristics of online social network data

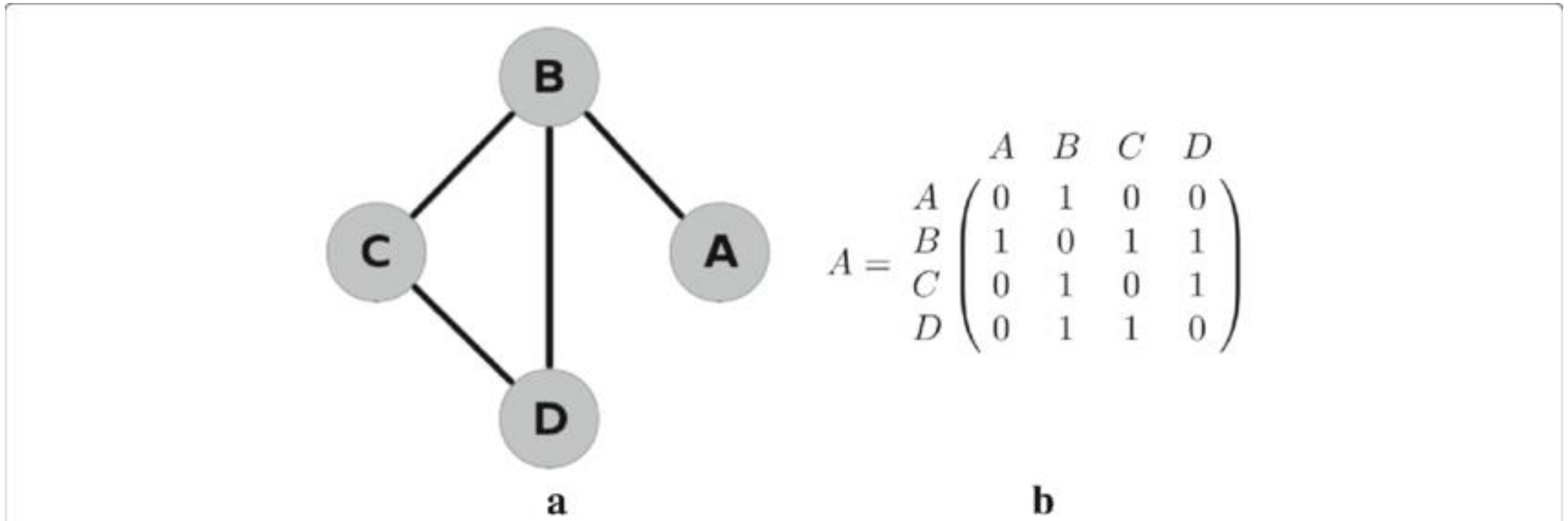
- Real-time view on society
- Different types of data together: text, location, timestamp, images
- Straightforward network structure with different types of interactions

Networks in a nutshell

Graph G(V,L): finite set V of *n* elements (nodes, vertexes) and set of *k* couples of nodes (links, edges)

A network/graph can be defined through its *adjacency matrix* -> compute centrality measures

e.g. connectivity degree k_i: row/column sum (number of neighbours of a node)



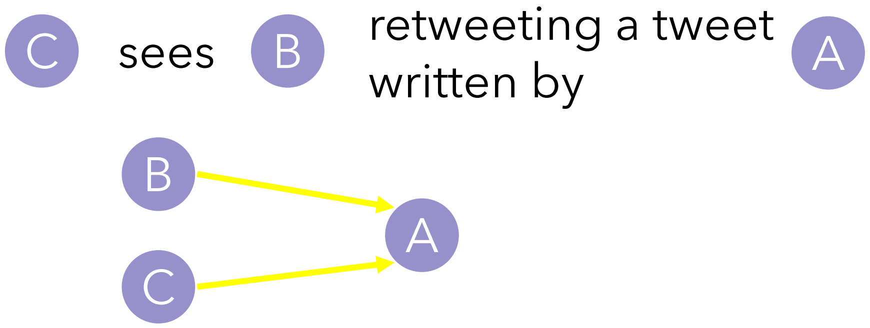
Retweet network (X platform)

A→B if user A **retweets** user B.

Weight: number of retweets.

Interpretation: agreement.

The edge is always between the retweeting user and the writing user: the intermediate retweet structure is hidden. One tweet form a star-like graph, so the final network is an aggregation of star-like modules. Densely connected modules can be thought of people sharing the same ideas.



Network analysis

- Centrality measures
- Community structure
- Linking network structure to other attributes: text, geo-localization

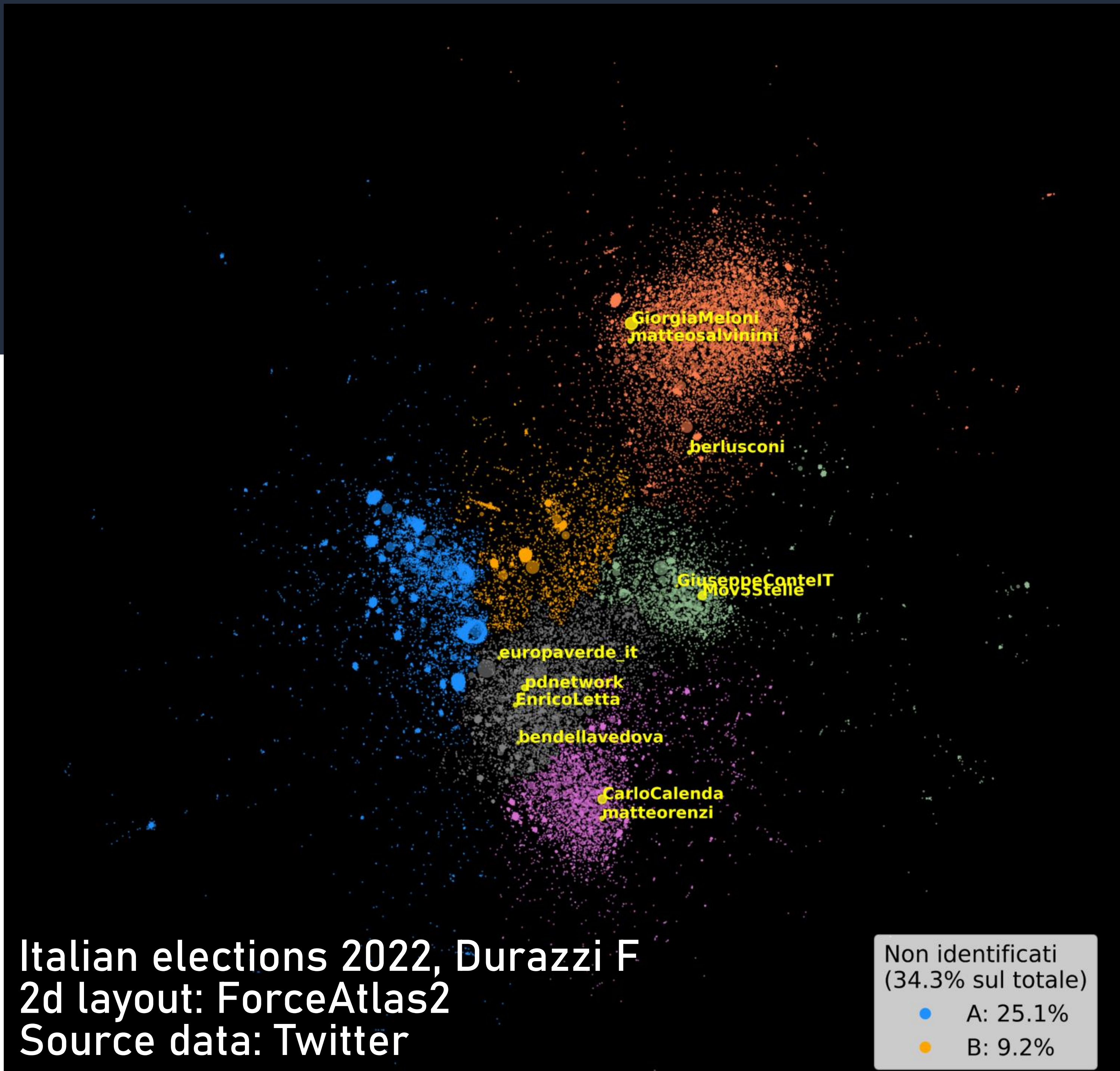
Community detection

Objective: find groups of users maximizing the intra-community modularity Q

$$Q_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

In plain words, maximize the number of edges inside the community w.r.t. those expected by chance

See more details on my Linkedin



Not just networks: text embeddings

Mapping text into vector spaces allows Machine Learning and Artificial Intelligence applications:

- Clustering
- Classification
- Regression
- Vector operations (eg sum/difference, average)

Vocabulary lookup table

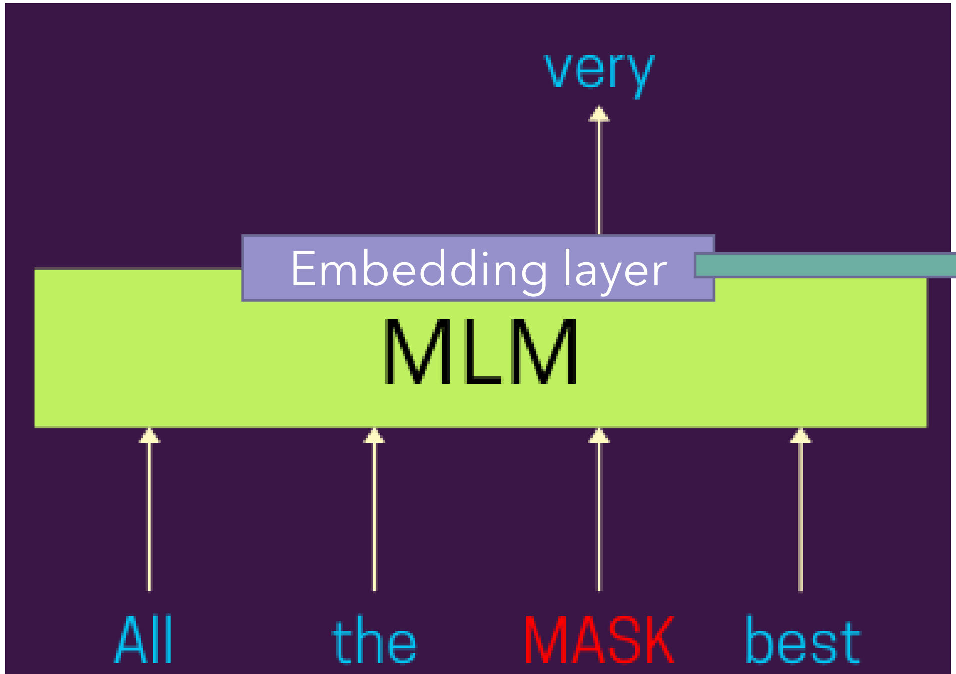
$$\text{I love chocolate.} \leftrightarrow \underbrace{\begin{matrix} 0.3 & -0.2 & 0.9 \\ 0.4 & 1.3 & -1.2 \\ 0.2 & 3.5 & -0.6 \\ 1.1 & -1.2 & 0.8 \end{matrix}}_{\text{\# of words}}$$

Embedding dimension

Not just networks: text embeddings

Language Modeling (LM) is a very popular way to embed text using neural networks (specifically *attention-based transformers*)

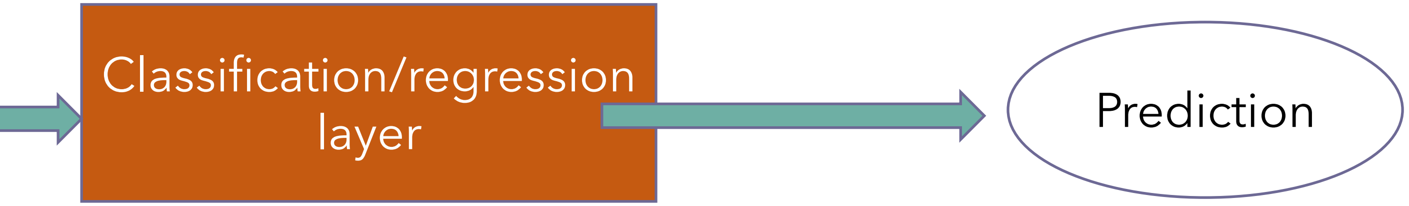
1) **Pre-training** as Language Model: model predicting missing/masked words in a sentence (self-supervised on large corpora, e.g. www, Wikipedia, Twitter) -> to predict the masked word, all the sentence is encoded in its embedding layer



2) Task-specific **fine-tuning**: final classification/regression layer for the final task (supervised regression/classification on labelled data)

With Step1, the models “learns the language” in general and with Step2, it learns how to deal with specific tasks

You can extract sentence-embeddings as the second-last layer (the one before masked-token prediction)



Vaswani A, *Attention is all you need*, 2017

https://huggingface.co/blog/getting-started-with-embeddings

Example: Twitter retweet network on «vaccination»

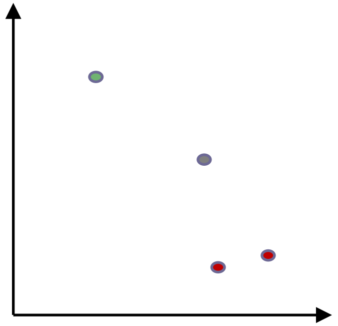
ProVax and AntiVax classification through network and text fetaures

- **Embedding of Tweets text** into a N-dimensional

Text	Label	User ID
Odio i vaccini	AntiVax	1
Non vaccinatevi mai	AntiVax	1
Oggi partono le vaccinazioni.	Neutral	2
Vaccino fatto	ProVax	2

Gori D, *Mis-tweeting communication: a Vaccine Hesitancy analysis among twitter users in Italy*, Acta Biomedica, 2021

Text embeddings



Example: Twitter retweet network on «vaccination»

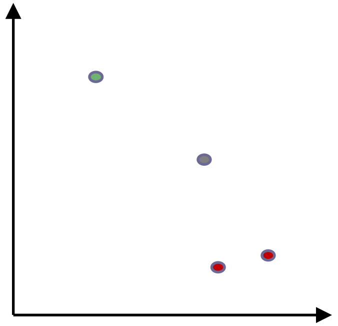
ProVax and AntiVax classification through network and text fetaures

- **Embedding of Tweets text** into a N-dimensional space (BERT transformer)
- Represent **users as community-based vectors** (participation ratio)

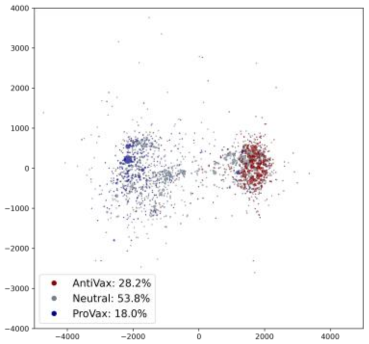
Text	Label	User ID
Odio i vaccini	AntiVax	1
Non vaccinatevi mai	AntiVax	1
Oggi partono le vaccinazioni.	Neutral	2
Vaccino fatto	ProVax	2

Gori D, *Mis-tweeting communication: a Vaccine Hesitancy analysis among twitter users in Italy*, Acta Biomedica, 2021

Text embeddings



Network 2d layout



Example: Twitter retweet network on «vaccination»

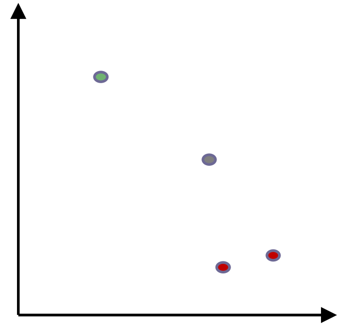
ProVax and AntiVax classification through network and text fetaures

- **Embedding of Tweets text** into a N-dimensional space (BERT transformer)
- Represent **users as community-based vectors** (participation ratio)

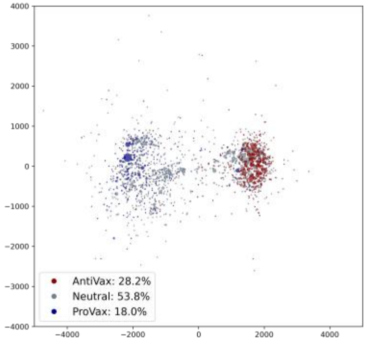
Text	Label	User ID
Odio i vaccini	AntiVax	1
Non vaccinatevi mai	AntiVax	1
Oggi partono le vaccinazioni.	Neutral	2
Vaccino fatto	ProVax	2

Gori D, *Mis-tweeting communication: a Vaccine Hesitancy analysis among twitter users in Italy*, Acta Biomedica, 2021

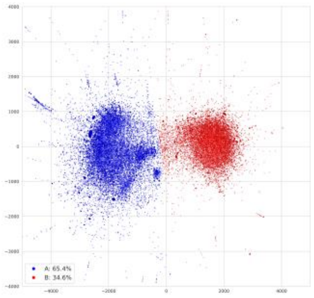
Text embeddings



Network 2d layout



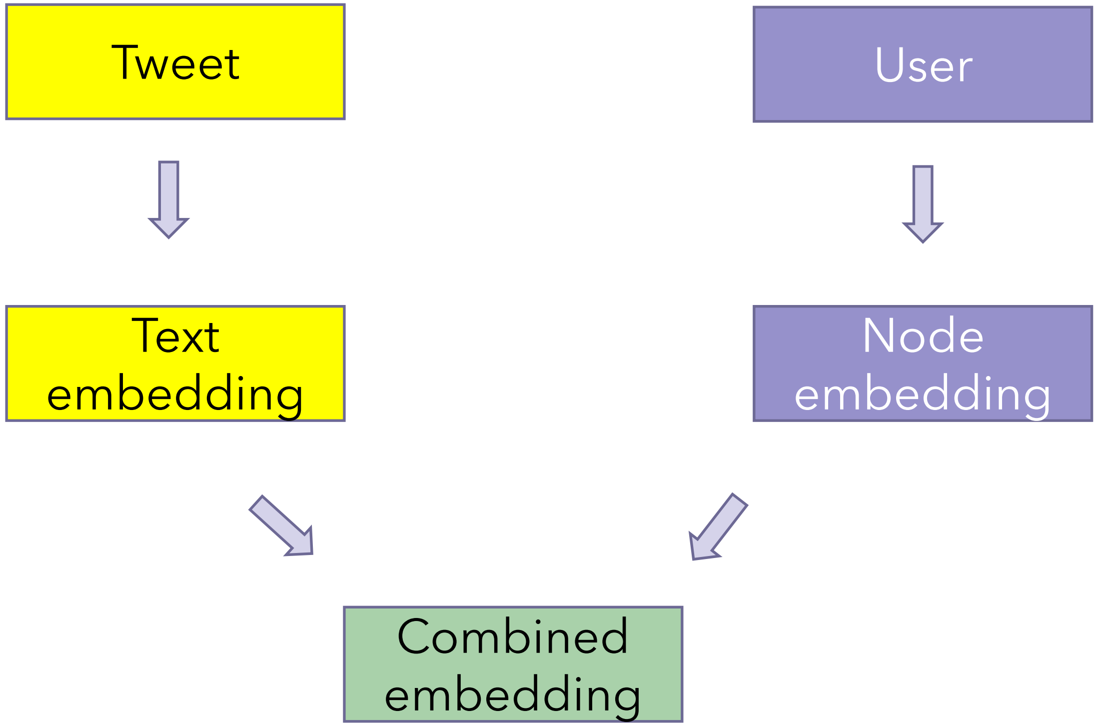
Predictions



Example: Twitter retweet network on «vaccination»

ProVax and AntiVax classification through network and text fetaures

- **Embedding of Tweets text** into a N-dimensional space (BERT transformer)
- Represent **users as community-based vectors** (participation ratio)
- Merge text and network features to classify users (Deep learning or simpler)
- **Merge text and network features to classify users** (Deep learning or simpler)



Scientific literature automated search and analysis

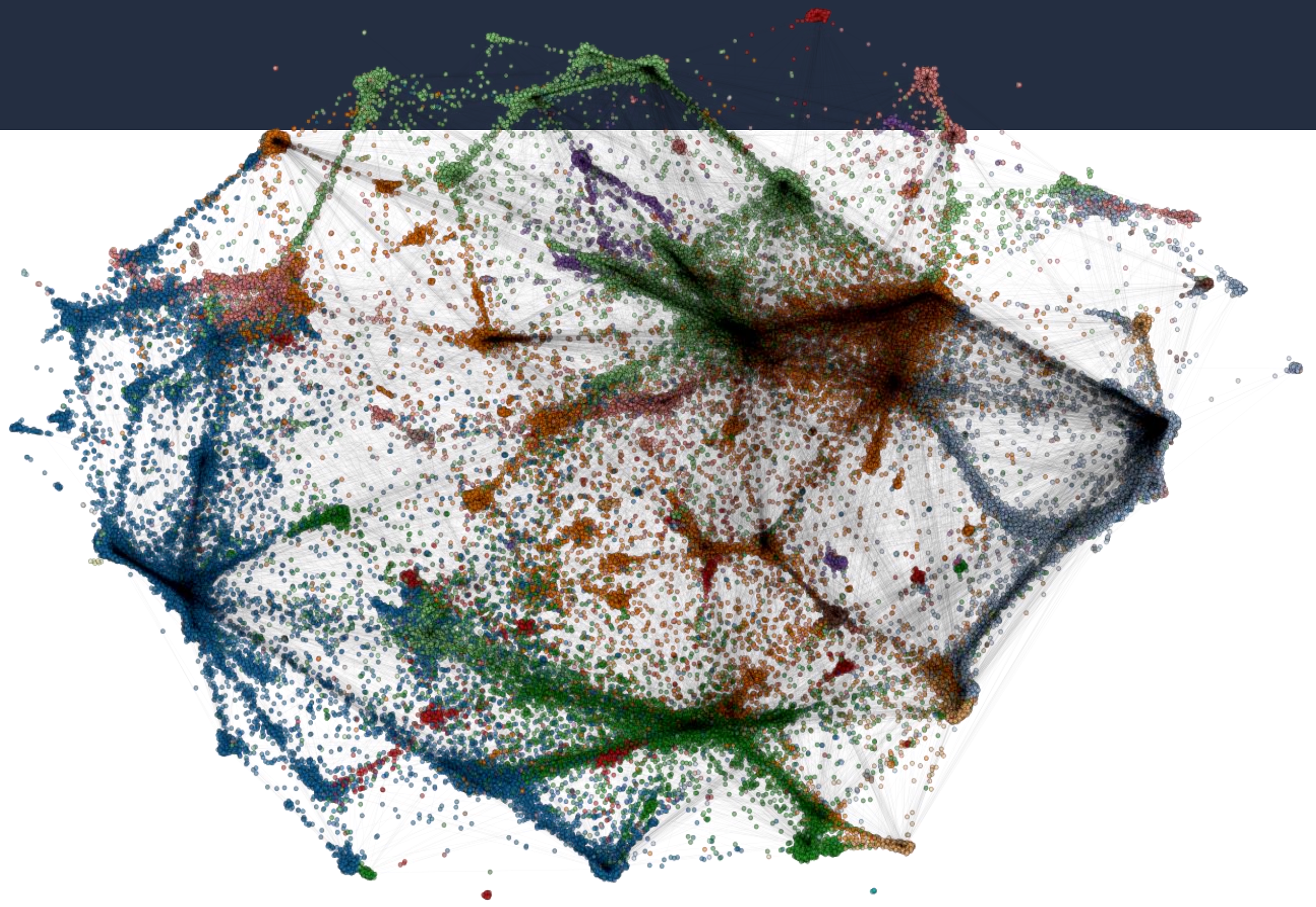
We collected the whole Pubmed archive

Choose topics of interest: e.g. COVID-19

Citation network

Topic modelling: transform abstracts into vectors and clusterize them to extract the different topics

Natural Language Processing and Regular Expressions to extract information: values, keywords, results



Scientific literature automated search and analysis

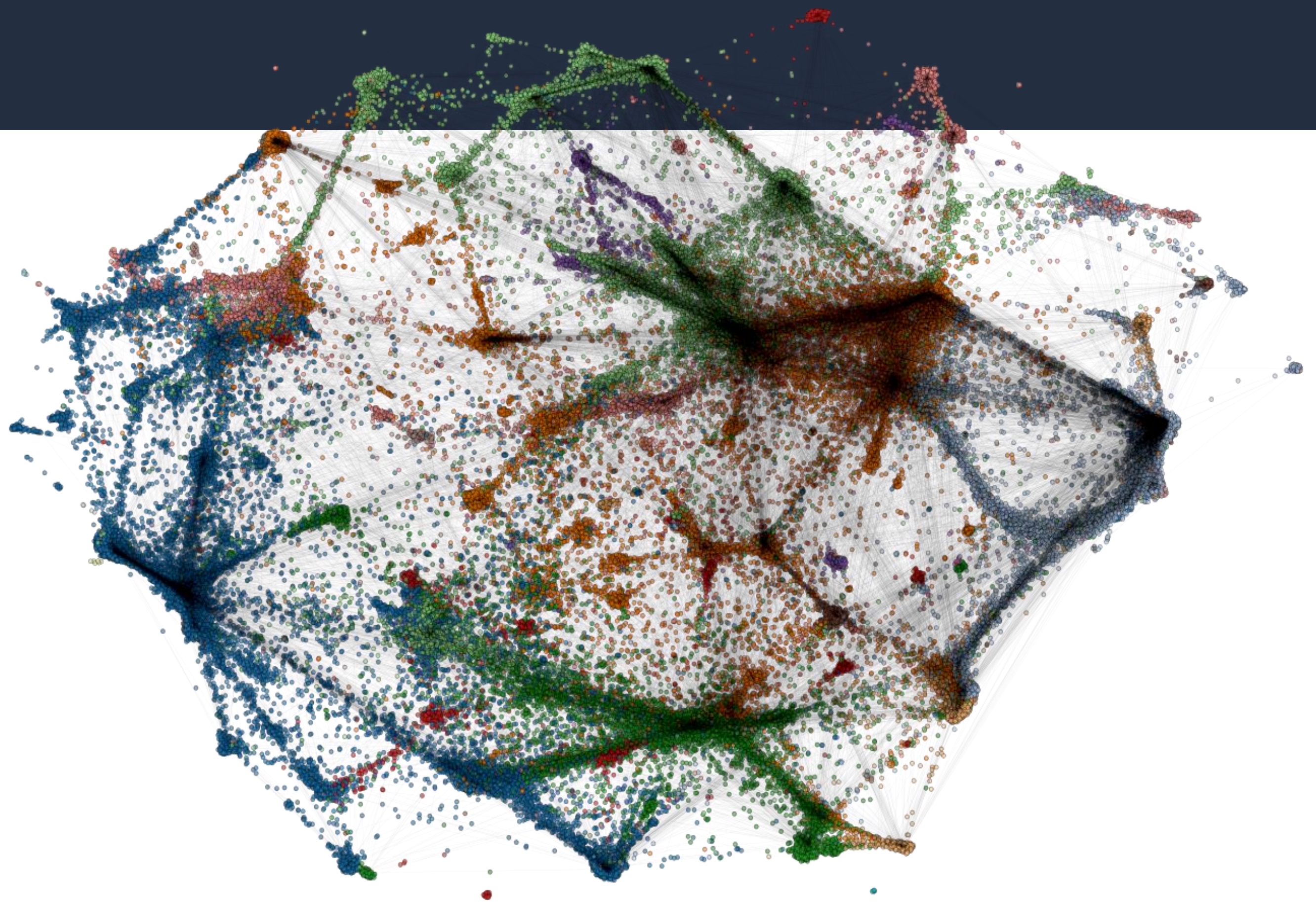
Early detection of relevant papers

Automated approach to:

- Measure citation growth speed

- Build autor-level features

- Explore the relationship between groundbreaking papers and network structure: "hub" authors & papers



Success is driven by connections?
https://bigthink.com/the-well/the-science-of-success/

Citations are indicators of good quality?

https://link.springer.com/article/10.1007/s11192-023-04735-0

Environmental traces

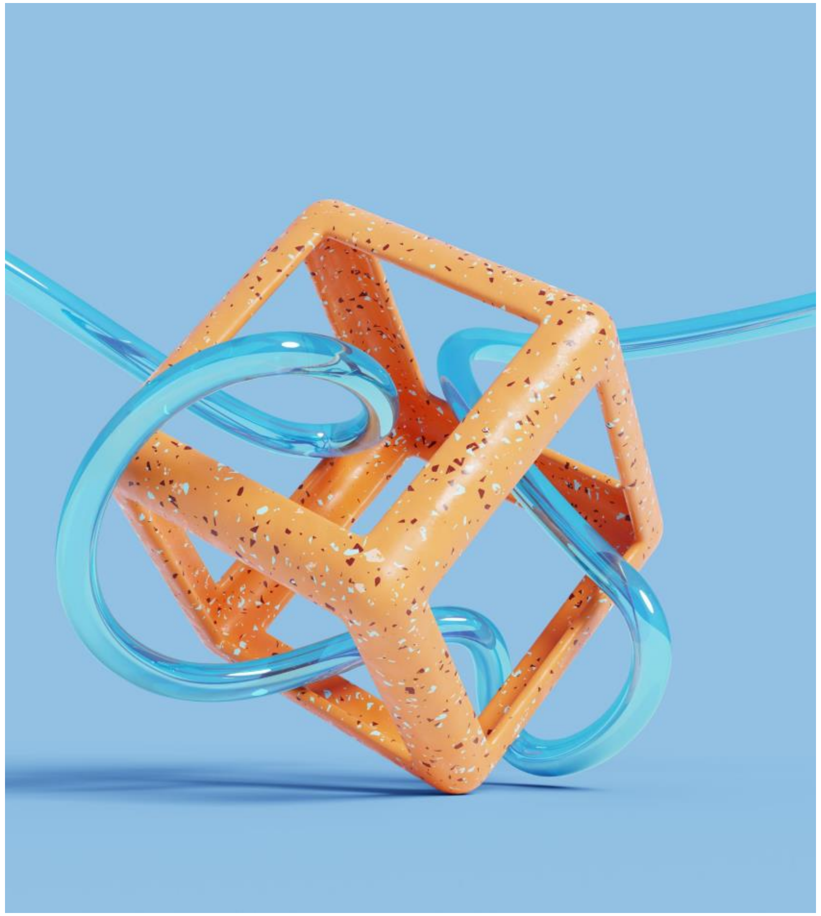
Epidemiological monitoring at urban level

Clinical and mobility data integrated to wastewater sequencing

Environmental traces

Epidemiological monitoring at urban level

Clinical and mobility data integrated to wastewater sequencing



Epidemic entanglement

Difficult to disentangle single contributions of epidemic drivers

Example: at now, the number of COVID-19-infected individuals is way lower than during the pandemic peaks of 2020-2021. This is due to:

a) lower transmissibility of the virus?
b) increased vaccination coverage?
c) mutated social habits? (distancing, facial masks)
d) different climatic conditions?
e) less testing

3-year monitoring of COVID-19 in Bologna metropolitan area

- Epidemiological mathematical model adjusted on clinical data
- RNA sequencing on urban sewage
- Emergence of SARS-CoV-2 lineages over time through genomic data
- Road traffic time series
- Vaccination coverage of the population

HERA

RNA sequencing of wastewater

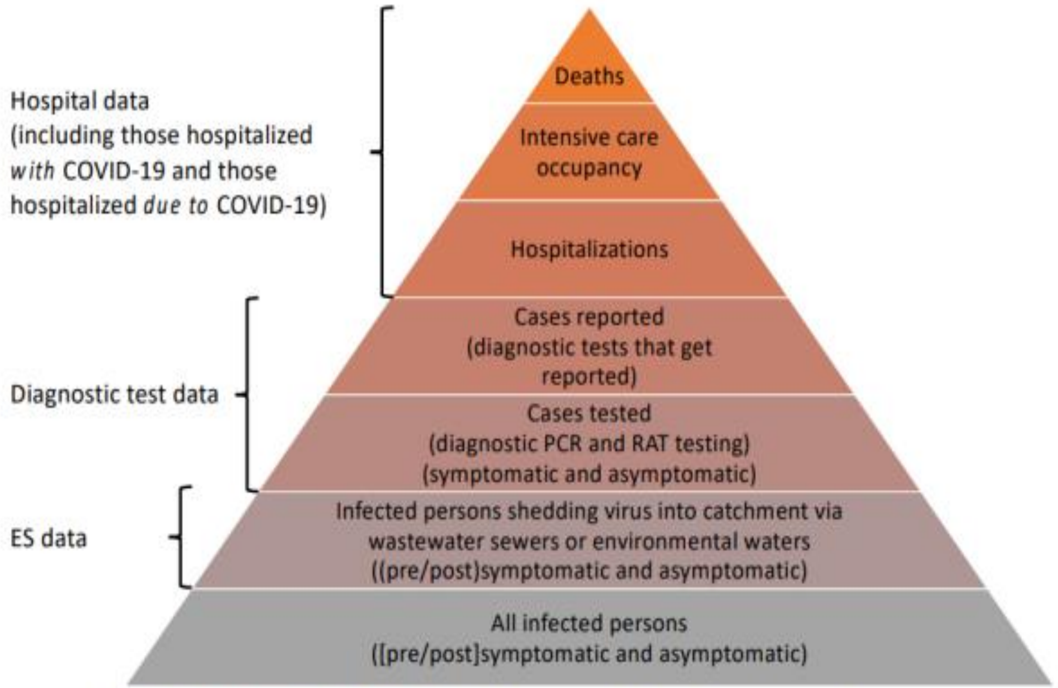
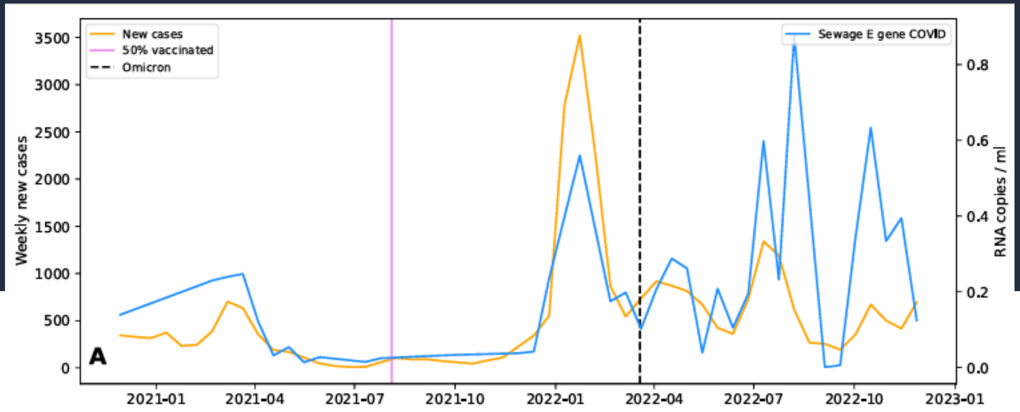


Figure 1. Illustration of the role of SARS-COV-2 environmental surveillance as a source of data on COVID-19 and SARS-CoV-2 in communities via a defined wastewater catchment[5].

[5] World Health Organisation (WHO) Environmental surveillance for SARS-COV-2 to complement public health surveillance, 14 April 2022. WHO-HEP-ECH-WSH-2022.1-eng.pdf. Environmental surveillance for SARS-COV-2 to complement public health surveillance – Interim Guidance (who.int)

- Martin, J.; Klapsa, D.; Wilton, T.; Zambon, M.; Bentley, E.; Bujaki, E.; Fritzsche, M.; Mate, R.; Majumdar, M. **Tracking SARS-CoV-2 in Sewage: Evidence of Changes in Virus Variant Predominance during COVID-19 Pandemic**. *Viruses* 2020, *12*, 1144. doi: 10.3390/v12101144
- Izquierdo-Lara R, Elsinga G, Heijnen L, Munnink BBO, Schapendonk CME, Nieuwenhuijse D, Kon M, Lu L, Aarestrup FM, Lycett S, Medema G, Koopmans MPG, de Graaf M. **Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium.** *Emerg Infect Dis.* 2021 May;27(5):1405-1415. doi: 10.3201/eid2705.204410.
- Nattino G, Castiglioni S, Cereda D, et al. **Association Between SARS-CoV-2 Viral Load in Wastewater and Reported Cases, Hospitalizations, and Vaccinations in Milan**, March 2020 to November 2021. *JAMA.* 2022;327(19):1922–1924. doi:10.1001/jama.2022.4908

RNA sequencing of wastewater



Methods

- Sampling twice per month (November 2020 – November 2022)
- RNA extraction and real-time PCR on SARS-CoV-2 E-gene
- Viral load estimation from serial dilutions

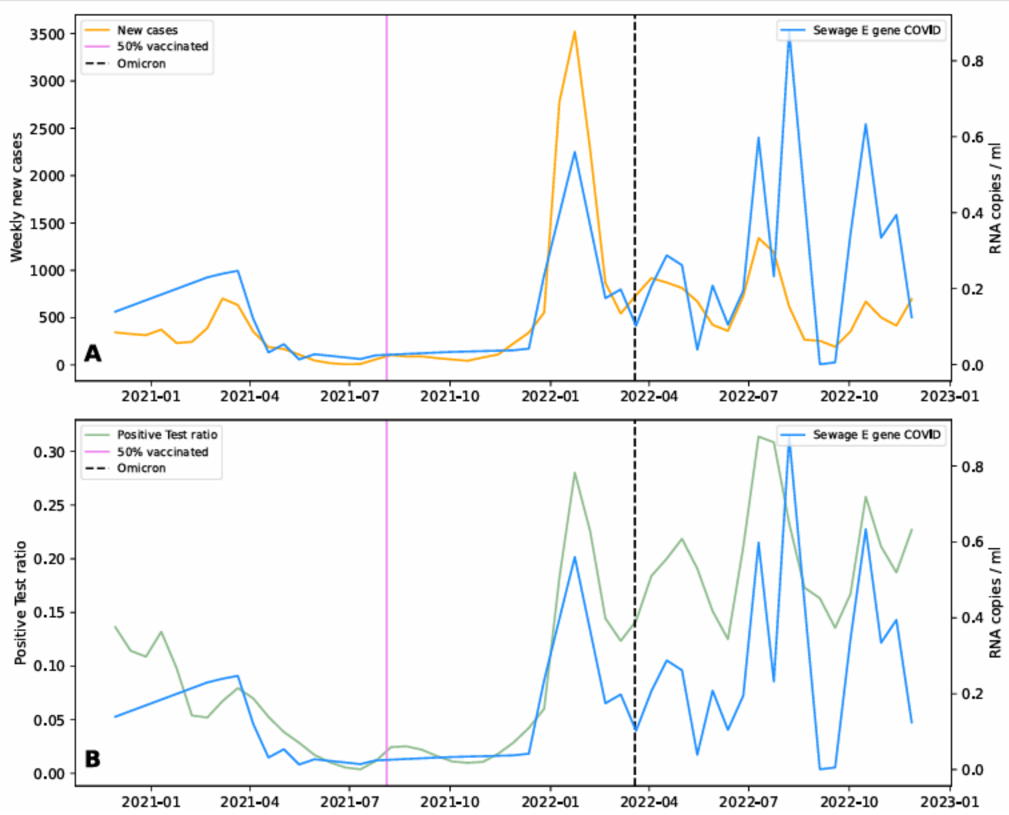
Results

- Correlation between sewage viral load and number of cases: positive test ratio (r=0.73)
- Hospitalizations decline but viral load increase

RNA sequencing of wastewater

Methods

- Sampling twice per month (November 2020 – November 2022)
- RNA extraction and real-time PCR on SARS-CoV-2 E-gene
- Viral load estimation from serial dilutions



Results

- Correlation between sewage viral load and number of cases: positive test ratio (r=0.73)
- Hospitalizations decline but viral load increase

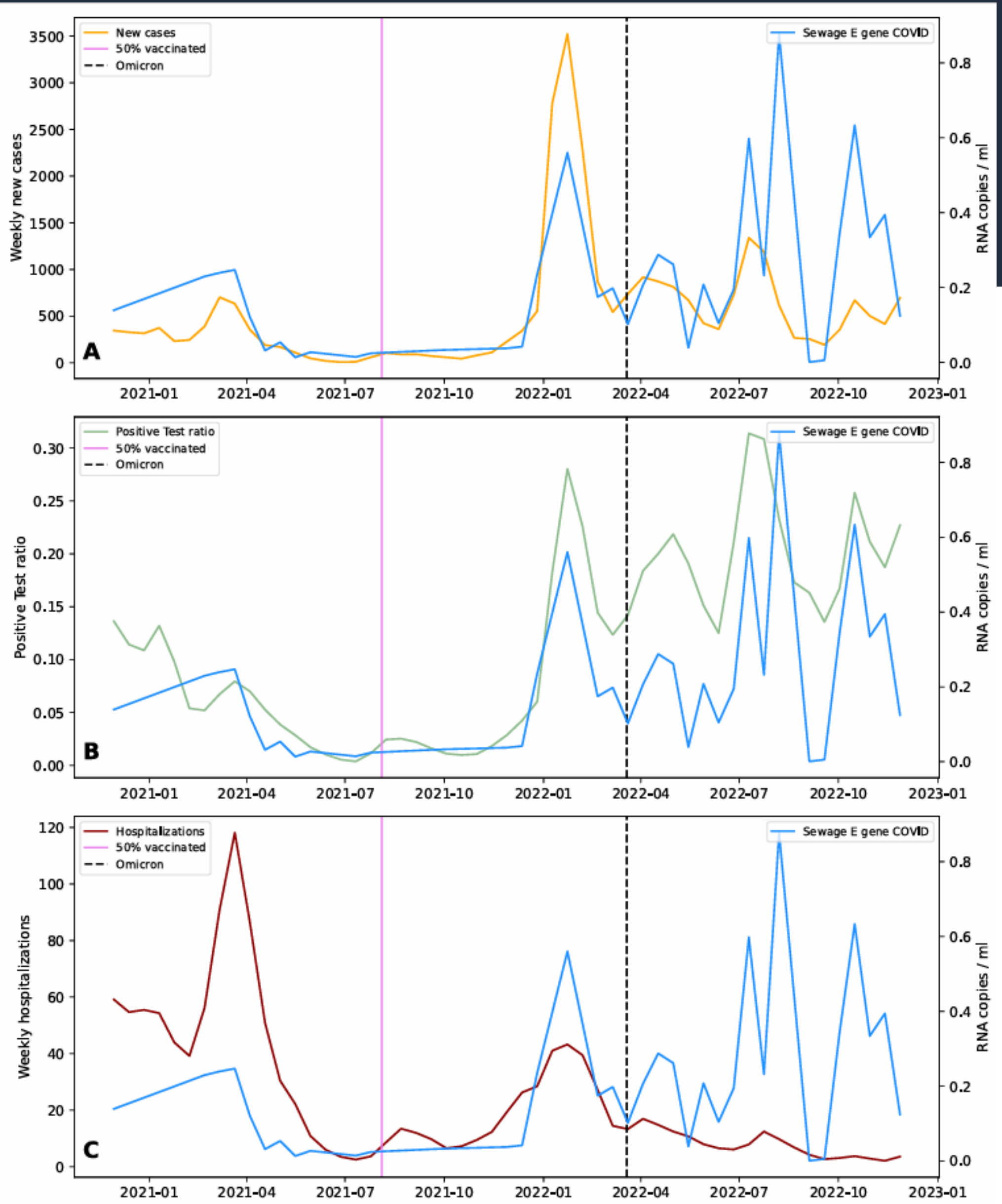
RNA sequencing of wastewater

Methods

- Sampling twice per month (November 2020 – November 2022)
- RNA extraction and real-time PCR on SARS-CoV-2 E-gene
- Viral load estimation from serial dilutions

Results

- Correlation between sewage viral load and number of cases: positive test ratio (r=0.73)
- Hospitalizations decline but viral load increase



Sociability and mobility

Sociability: amount of social activity, estimated from the number of infections through an epidemiological model

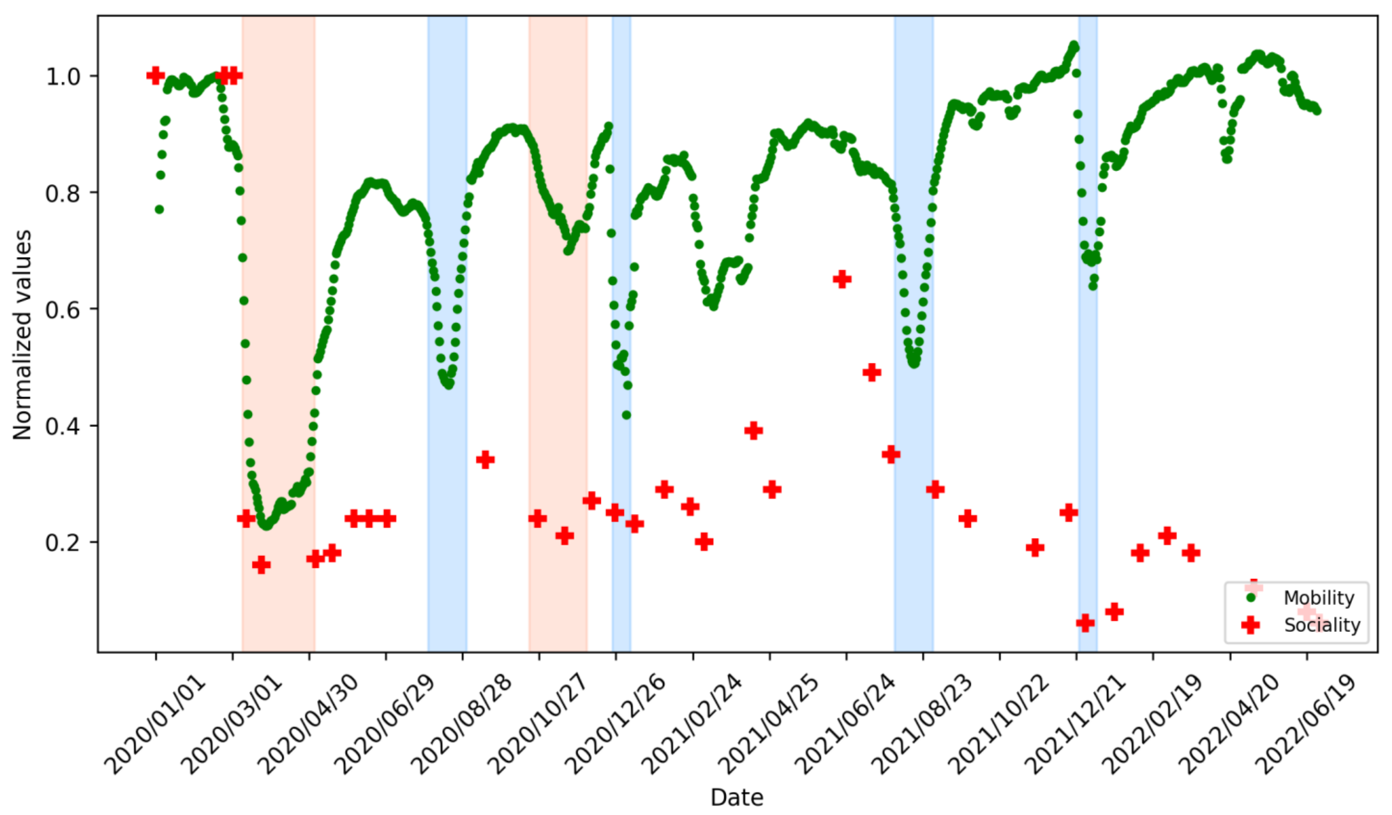
Mobility: measured from road traffic in Bologna

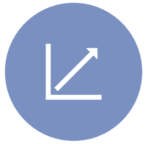
Sociability and mobility

Sociability: amount of social activity, estimated from the number of infections through an epidemiological model

Mobility: measured from road traffic in Bologna

- **Red areas**: lockdowns and curfews
- **Blue areas**: holidays



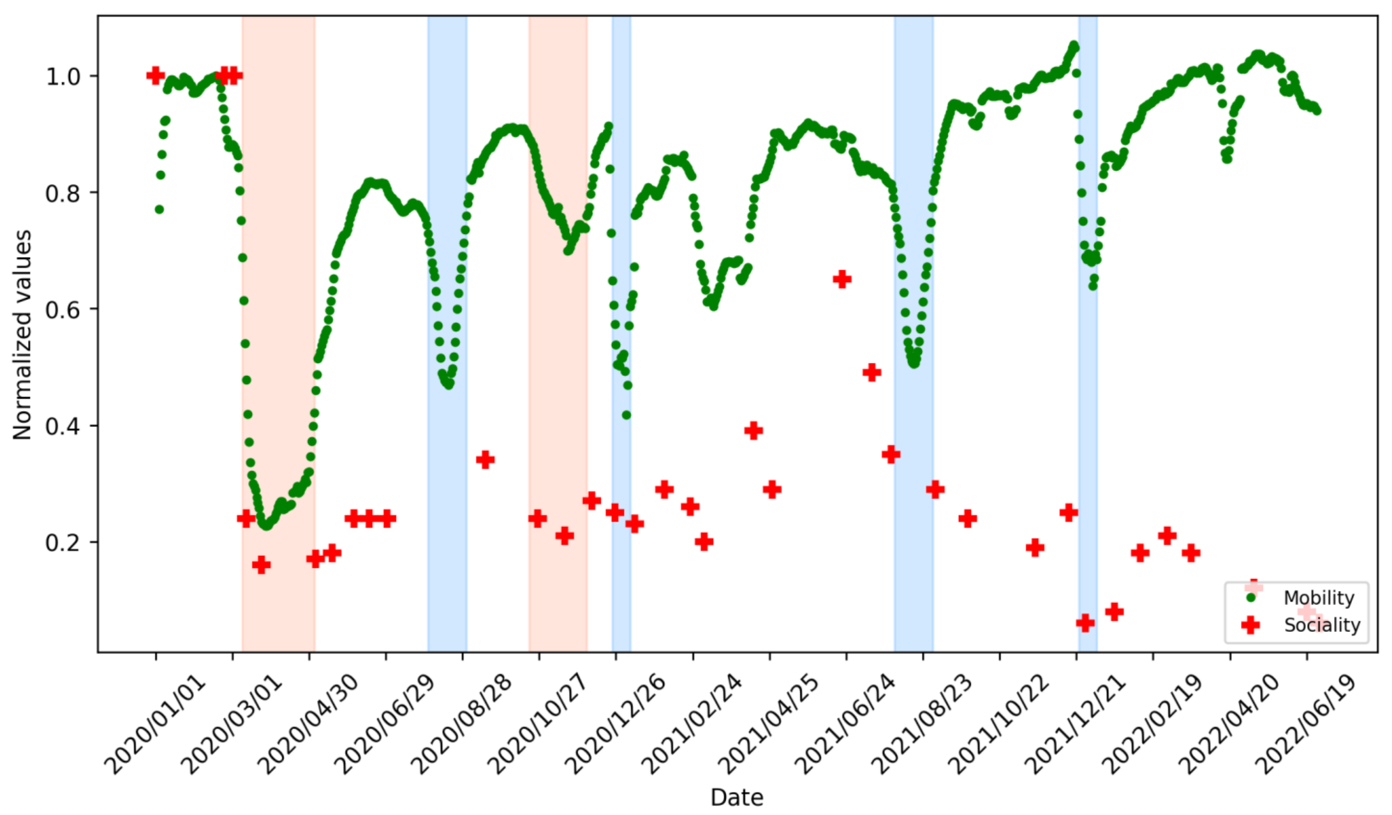


Sociability and mobility

Sociability: amount of social activity, estimated from the number of infections through an epidemiological model

Mobility: measured from road traffic in Bologna

- **Red areas**: lockdowns and curfews
- **Blue areas**: holidays



- **Mobility** is critically impacted at the first lockdown (February 2020)
- **Mobility** slowly recovers to pre-pandemic values, with down-ward peaks at holidays and closures

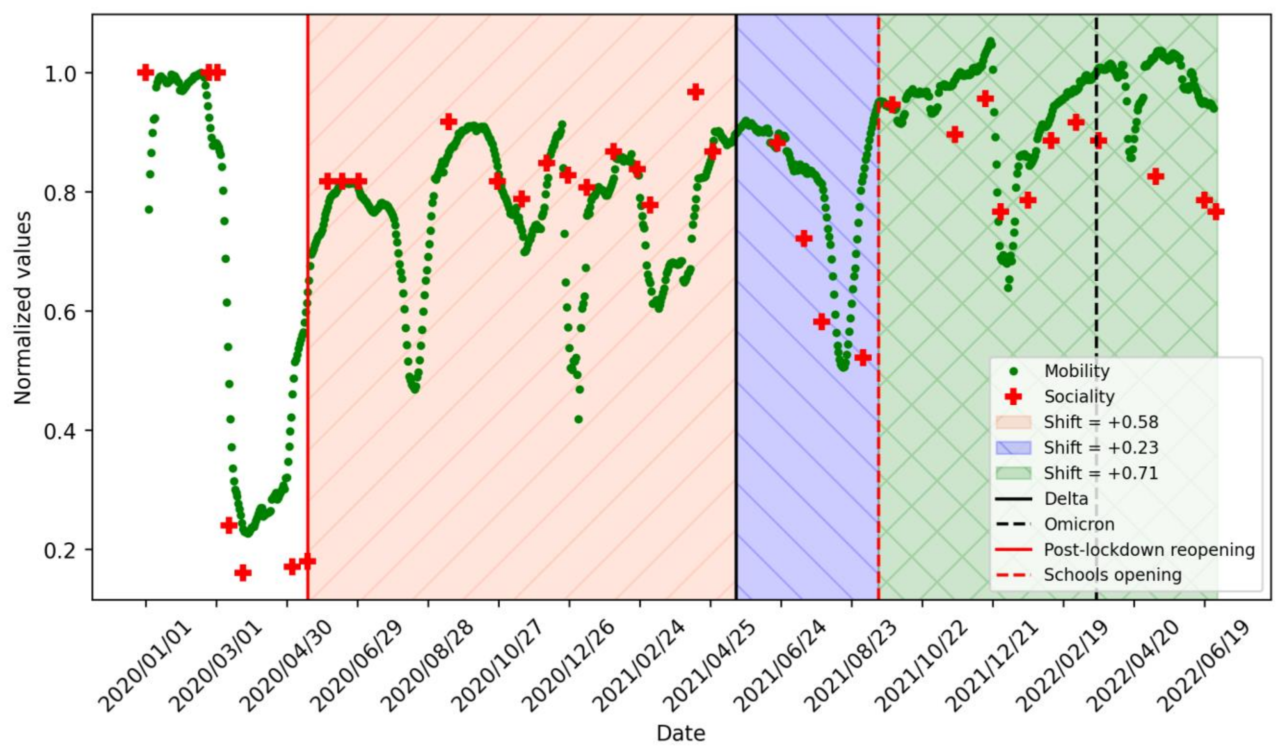
- **Sociability** impacted at the first lockdown (same as mobility), but remains generally low -> contacts are now protected and viruses are weaker

Sociability and mobility

3 breakpoints: shift to re-normalize mobility and sociability (on the first value)

Day	Event	Shift
18/05/2020	Activities reopening (bar, restaurants)	0.58
17/05/2021	Delta variant in Emilia Romagna	0.23
15/09/2021	Schools reopening	0.71

- High correlation (r=0.76)
- **Mobility can be used as a proxy to parametrize sociability** in the model for *short* periods (3 months at least)



- **Shifts ~ gap between protected and unprotected contacts**
- Small shift at outbreak (still not much protection) and summer 2021
- Larger shifts during periods of increased sensitivity to control measures (distancing, facial masks)

Conclusions

We live in an era where many traces are available:

to big tech corporates: surveillance capitalism ☹

to «investigative» scientists: reveal unexpected associations and hidden correlations ☻

For a physicist, new areas emerge in which laws can be proposed and their validity verified through measurements and experiments

Conclusions

We live in an era where many traces are available:

to big tech corporates: surveillance capitalism ☹

to «investigative» scientists: reveal unexpected associations and hidden correlations ☻

For a physicist, new areas emerge in which laws can be proposed and their validity verified through measurements and experiments

Happy hunting for traces!