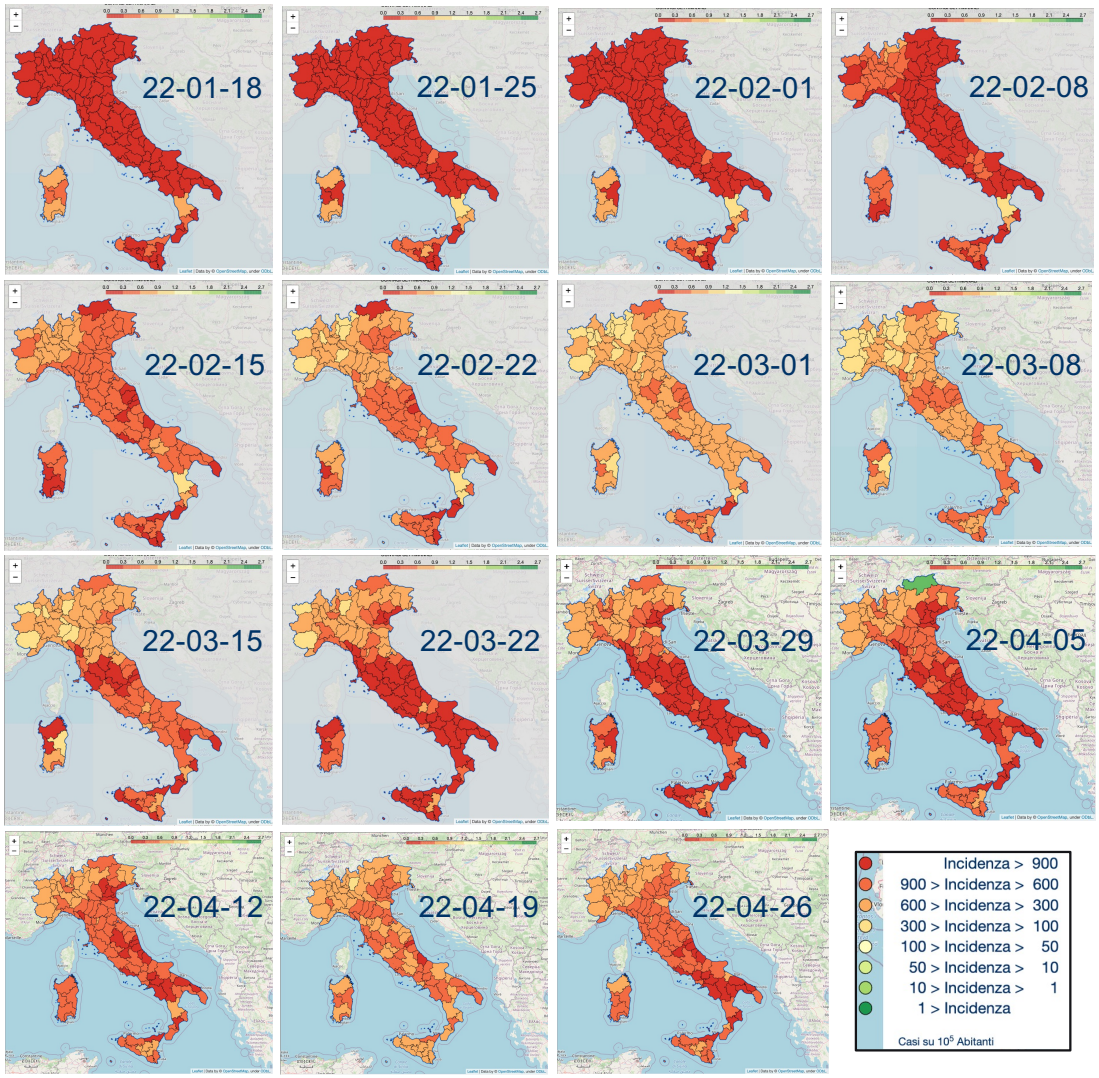
Experience with Statistical Analysis of Covid data

Gaetano Salina
INFN Sezione di Roma Tor Vergata



Summary:

- *Data Definition, Data Acquisition & Computational Models*

Is it a well-defined problem ? Maybe not !

- *Epidemic dynamics on Graphs*

Epidemic Models & Complex Systems: the role of topology

- *Experience with Covid Data Analysis*

Creating a Knowledge Base for Covid Data: a Nightmare

I apologize for my broken English and for the large number of slides

The new gospel:

With enough data, the numbers speak for themselves.

Chris Anderson
An English-American author and entrepreneur.
He was the WIRED editor-in-chief until 2012.

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

"All models are wrong, but some are useful."

This article has been reproduced in a new format and may be missing content or contain faulty links. Contact wiredlabs@wired.com to report an issue.

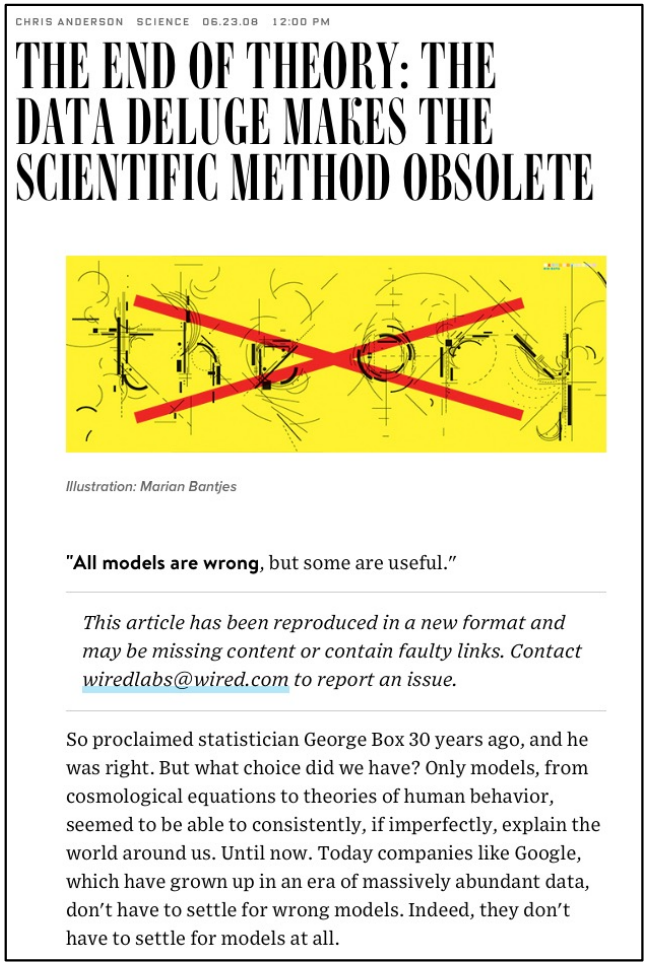
So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

https://www.wired.com/2008/06/pb-theory/

The temptation: *Correlation is enough !*

- Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."
- Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.
- We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Chris Anderson
An English-American author and entrepreneur.
He was the WIRED editor-in-chief until 2012.



CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

"All models are wrong, but some are useful."

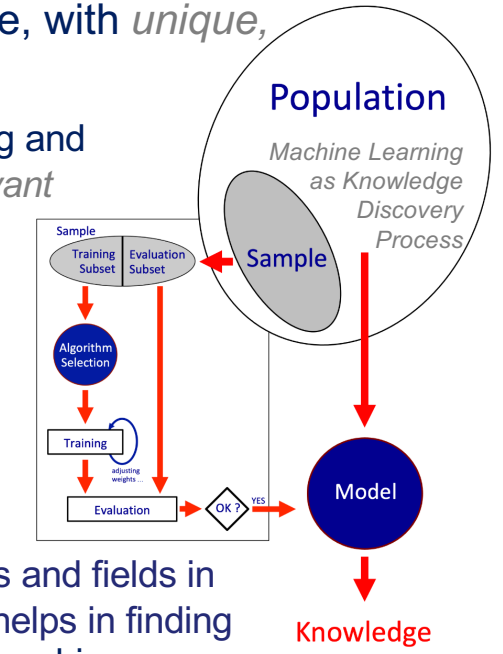
This article has been reproduced in a new format and may be missing content or contain faulty links. Contact wiredlabs@wired.com to report an issue.

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

https://www.wired.com/2008/06/pb-theory/

The Knowledge Discovery Process (KDP)

The Knowledge Discovery Process has the aim of providing decisions based on data. The ETL (Extract, Transform and Load, i.e. steps 1-3) is important to have an integrated Knowledge Base, with *unique, well-defined and quality data*, without redundancy.



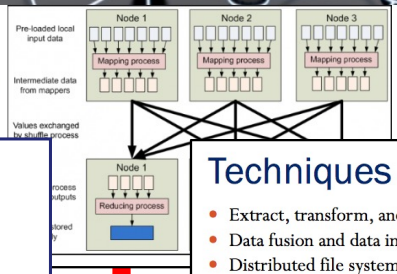
- Selection: Capturing relevant prior knowledge, identifying the data-mining goal and developing and understanding of the application domain. Based on that, *proper data samples as well as relevant variables* can be selected.
- Pre-processing: The selected data are processed. Handling of missing values, the noise and errors identification (and correction), the elimination of duplicates, as well as the matching, fusion, and conflict resolution for data taken from different sources are done.
- Transformation: The clean dataset is transformed into a form suitable for data mining algorithms analysis. To improve the analysis performance dimensionality reduction methods can also be applied.
- Data mining: Goals of data mining are prediction and description. In prediction some variables and fields in the dataset are used to predict unknown values of other variables of interest,and description helps in finding human-understandable patterns describing the data. Or: it involves the use of statistical and machine learning algorithms to extract structures, correlations, patterns and rules within data.
- Evaluation and interpretation: The patterns and models derived are analysed with respect to their validity. The user assesses the usefulness of the found knowledge. A data visualization of the extracted patterns and models is involved.

The KDP: some questions !

Knowledge

Distribution of resources and services

- Distributed Architecture
  - Clusters of computers that work together to a common goal
  - Scale out not up!
- Fault- tolerance
  - Resource replication
  - Eventual consistency
- Distributed processing
  - Shared-nothing model
  - New programming paradigms
- Quantum Computers



We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

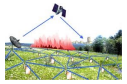
Chris Anderson

Real World

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

Data (Big)

Techniques for(big)data analysis

- Extract, transform, and load (ETL)
- Data fusion and data integration
- Distributed file system
- NoSQL database systems
- Cloud computing
- Analytics
  - Data mining
    - Association rule learning
    - Classification
    - Cluster analysis
    - Regression
  - Machine learning
    - Supervised learning
    - Unsupervised learning
- Crowdsourcing
- …

Computational Tools

Is it really so?

The KDP: some questions !

We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Chris Anderson

The(big)data process

Goal:
to make effective strategic decisions exploiting the availability of big data

When is a KDP successful?

My business has increased by 10%

E-Promotions:
Based on your current location, on your purchase history, on what you like ➔ the store located next to you sends promotions right now to you.

Only 10% of the abnormal situations are detected

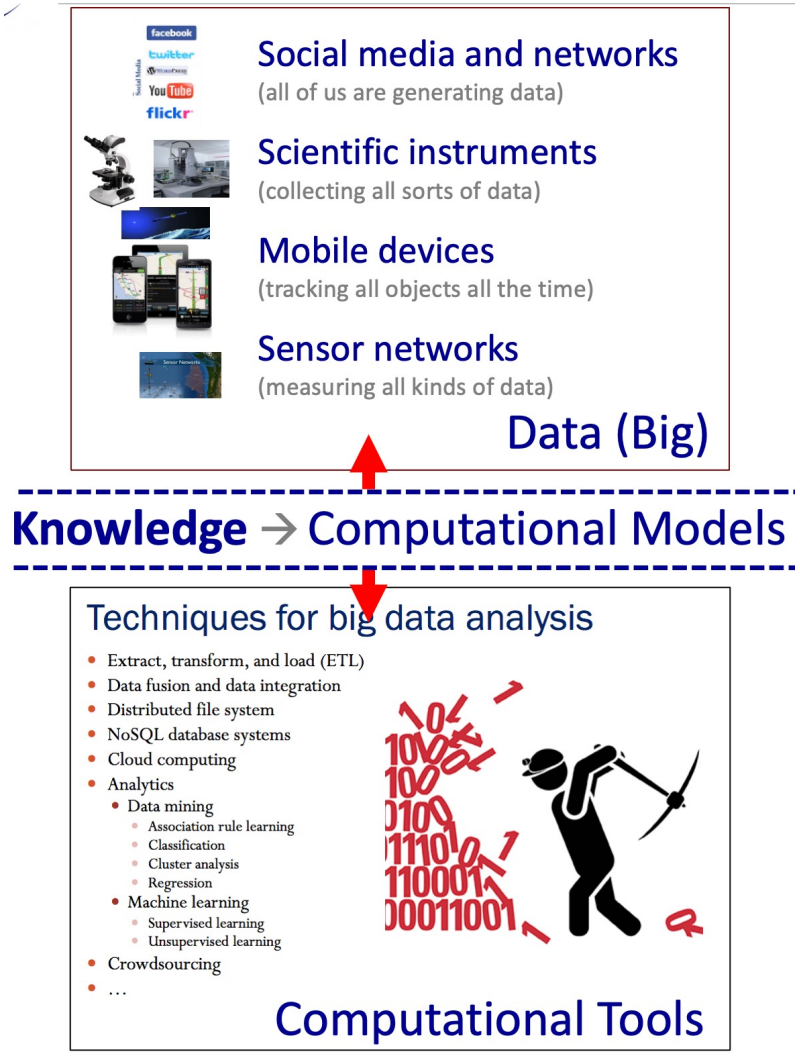
Healthcare monitoring:
Sensors monitoring your activities and body ➔ any abnormal measure requires immediate reaction.

Extracting Knowledge means developing Computational Models

Computational models are mathematical models that are simulated using computation to study complex systems. … The parameters of the mathematical model are adjusted using computer simulation to study different possible outcomes.

www.nature.com/subjects/computational-models

This may be a restrictive definition



Extracting Knowledge means developing Computational Models

Knowledge → Computational Model means:

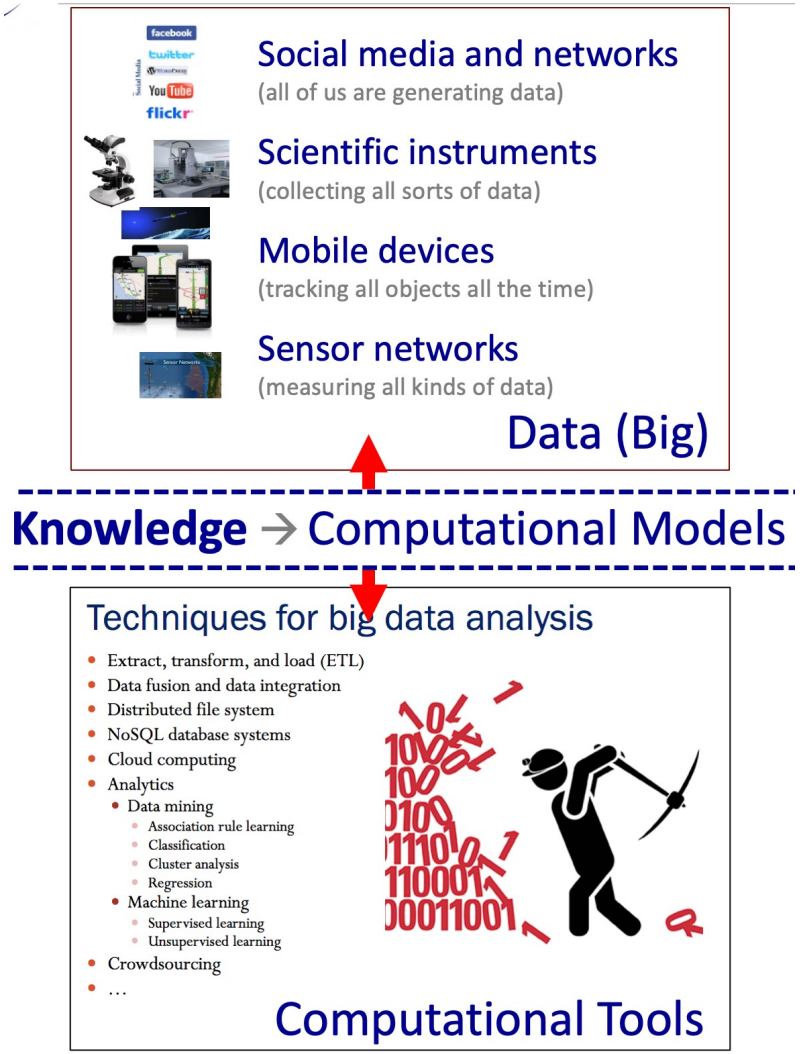
Knowledge → Theory → Algorithms.

We know the relevant observables for describing a system and their interaction laws. Computational models allow us to solve system with a large number of degree of freedom and complex dynamic or with non linear and/or non local interaction. Computational models are used when analytical tools fail or are unsuitable.

or

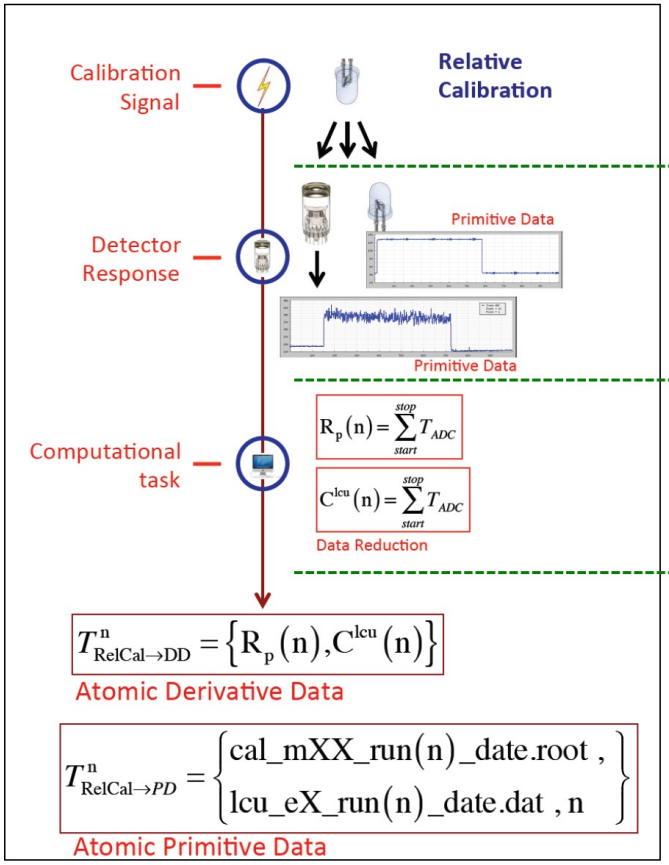
Knowledge → Heuristic → Algorithms.

No theory is available and/or we are not sure that the observables are relevant and/or the system is described by non-numerical observables. We use trial errors heuristic approach to define the observables interactions and define some numerical algorithms. Having a large number of degree of freedom and semantic different data we are playing in the field of Big Data. Computational models help us do define the correct observables and the system *dynamic.*



Extracting Knowledge means developing Computational Models

- Theoretical Models
- Measurement Theory
- Development of Detectors and Measuring Instruments



Physical Systems

Knowledge → Primitive Data

Knowledge → Derivative Data

Data

Computational Models

How Physics tamed and tames (Big) Data

Knowledge → Heuristic → Algorithms.

No theory is available and/or we are not sure that the observables are relevant and/or the system is described by non-numerical observables. We use trial errors heuristic approach to define the observables interactions and define some numerical algorithms. Having a large number of degree of freedom and semantic different data we are playing in the field of Big Data. Computational models help us do define the correct observables and the system *dynamic.*

Knowledge → Theory → Algorithms.

We know the relevant observables for describing a system and their interaction laws. Computational models allow us to solve systems with a large number of degree of freedom and complex dynamics or with non linear, non local interactions. Computational models are used when analytical tools fail or are unsuitable.

Theoretical Models

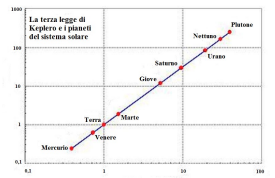
Measurement Theory

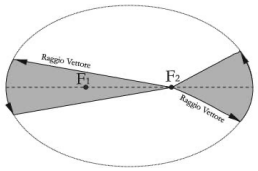
Development of Detectors and Measuring Instruments

Brahe was a towering figure. He ran a huge research program with a castle like observatory, a NASA-like budget, and the finest instruments and best assistants money could buy.

Brahe

http://www.chrisbaldassano.com/blog/2015/05/11/bigdata/



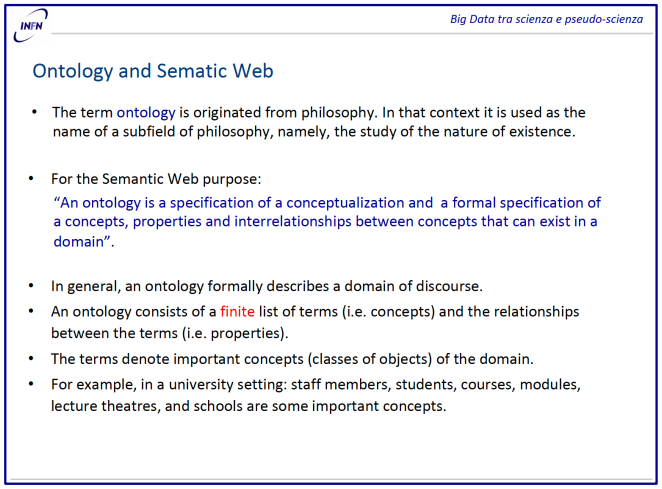


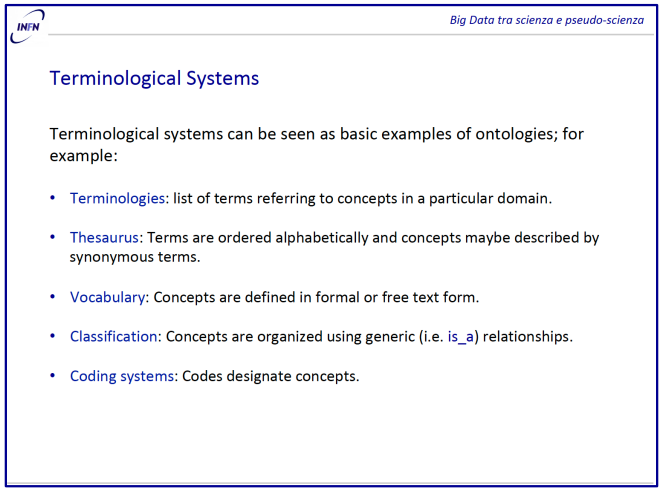
- The orbit of a planet is an ellipse with the Sun at one of the two foci.
- A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.
- The square of a planet's orbital period is proportional to the cube of the length of the semi-major axis of its orbit.

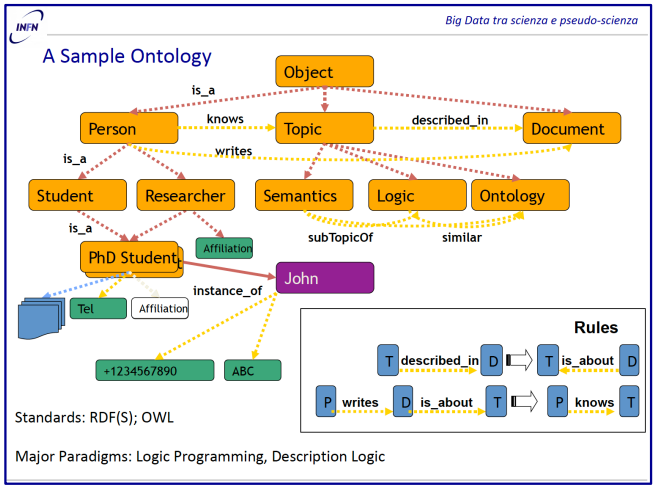
Keplero

A Data Ontology & Data Semantic is necessary

Semantics are only one part of the solution and often not the end-product so the focus of the design should be on creating effective methods, tools and APIs to handle and process the semantics. The concept of semantic data have changed the approach to data managing, due to the heterogeneous and unstructured nature of data sets. The ontological & semantic approach has been thought to manage data sets of this kind, i.e. a way to Knowledge Discovery Process.







It means developing Computational Models

The Knowledge Discovery Process (KDP)

Structure, Control and Computational Models

Building Energy Management Systems

Very general and high-level approach

Data Acquisition

MetaData

$$\left(y_1^{Type1}, y_2^{Type1}, \ldots, y_M^{TypeM_T}\right) \rightarrow M_D = \left\{y_i^k(t)\right\}_{M, M_T}$$

Detectors

$$\left(x_1^{Type1}, x_2^{Type1}, x_3^{Type2}, \ldots, x_N^{TypeN_T}\right) \rightarrow S = \left\{x_i^k(t)\right\}_{N, N_T}$$

Data Storage

DataBases

$$Stor\left(\left\{y_i^k\right\}\right) = \left\{M_D = \left\{y_i^k\right\}_{M, M_T}\right\}$$

$$Stor\left(\left\{x_i^k\right\}\right) = \left\{S = \left\{x_i^k\right\}_{N, N_T}\right\}$$

Data Visualization

Monitor & Analisys

$$Mon = \left\{M_\alpha\left(Stor\left\{x_i^k\right\}\right)\right\}$$

$$Mon = \left\{F_\beta\left(M_\alpha\left(Stor\left\{x_i^k\right\}\right)\right)\right\}$$

$$x_i^k \in S_A, S_B, \ldots, S_Z \subseteq \mathrm{S}$$

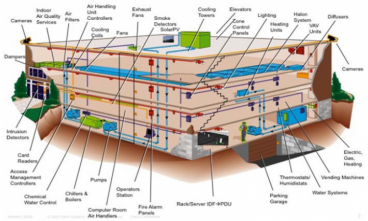
Data Optimization

Control & Optimization

$$F_C^\alpha = F\left\{M_\alpha, \left\{\lambda_k(t)\right\}, \left\{x_i^k\right\}, t\right\}$$

$$Min = \frac{\partial F_C^\alpha}{\partial\left\{\lambda_k(t)\right\}} = 0 \rightarrow \left\{\lambda_k^*(t)\right\}$$

Global Minimizations of Energy Functions



Data Path

Actuators

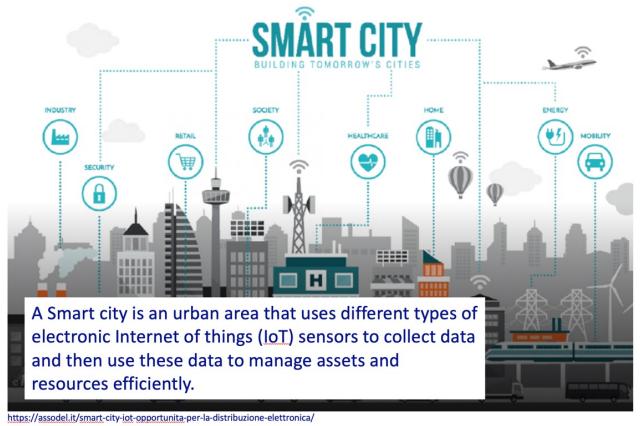
$$\left(w_1^{Type1}, w_3^{Type2}, \ldots, w_{N_w}^{TypeN_{T_w}}\right) \rightarrow W = \left\{w_i^k\left(\left\{\lambda_k(t)\right\}, t\right)\right\}_{N_w, N_{T_w}}$$

Data Path and Data Handling define the Computational Model of the system

$$ModComp = \left\{S, M_D, Stor, Mon, Min, W\right\}$$

The development of communication networks (wireless technology), low-cost and distributed computing power and Cloud-based software allow the creation of monitoring and control of systems with a high degree of freedom using data with different semantics and having evolution in space and time is determined by only partially known dynamics.

Smart Systems → Big Data.



https://assodel.it/smart-city-iot-opportunita-per-la-distribuzione-elettronica/

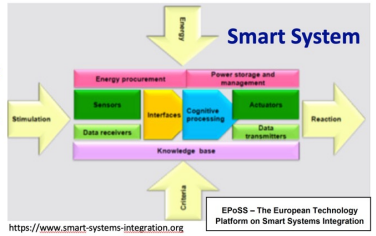
Smart Systems: Real systems that produce Big Data !

Smart Systems:

Real systems that produce Big Data !

Smart systems are self-sufficient intelligent systems with advanced functionality that provide solutions to address grand challenges and risks for mankind in social, economic and environmental terms.

They sense, diagnose, describe, qualify and manage a given situation in order to perceive complex circumstances, be predictive, and take autonomous decisions.



https://www.smart-systems-integration.org

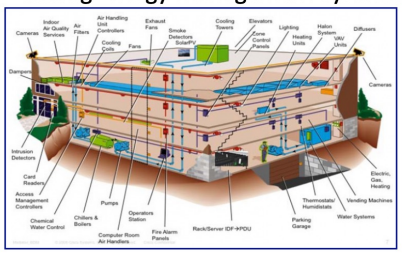
Smart Systems can be standalone, networked, or embedded into larger systems; they comprise heterogeneous devices combining data processing with sensing, actuating, energy scavenging, and communication and they excel in self-reliance and adaptability. What distinguishes smart systems from systems which are purely reactive is the **knowledge base**.

The Knowledge Discovery Process (KDP)

Structure, Control and Computational Models

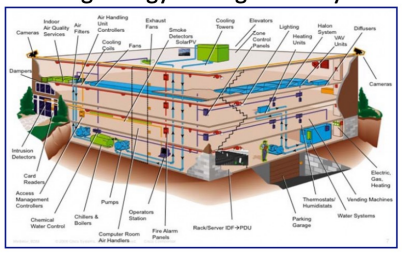
Computational Complexity: |ModComp|

Building Energy Management Systems



Computational Complexity is given by the physical system dynamics

Building Energy Management Systems



- High number and type of detectors and actuators
- The correlation length is of the order of the size of the system
- Complex Energy Function and complex optimization algorithms.

$\xi >> 1 \leftrightarrow \xi \sim 1$

- High number and type of detectors and actuators
- The System can be divided in non-interacting, simple, sub-systems.
- Simple Energy Function and well known optimization algorithms.

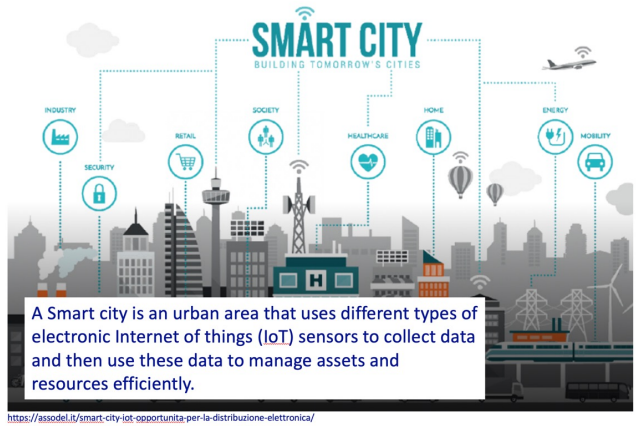
$$|ModComp|_{SBA} >> |ModComp|_{SBB}$$

Low Computational complexity systems do not require smart tools

Smart Systems require Computational Power

Gaetano Salina

Lecce 21 December 2021



https://assodel.it/smart-city-iot-opportunita-per-la-distribuzione-elettronica/

Smart Systems: Real systems that produce Big Data !

Gaetano Salina

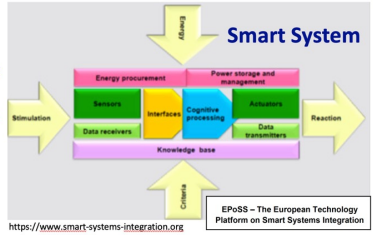
Lecce 21 December 2021

Smart Systems:

Real systems that produce Big Data !

Smart systems are self-sufficient intelligent systems with advanced functionality that provide solutions to address grand challenges and risks for mankind in social, economic and environmental terms.

They sense, diagnose, describe, qualify and manage a given situation in order to perceive complex circumstances, be predictive, and take autonomous decisions.



https://www.smart-systems-integration.org

Smart Systems can be standalone, networked, or embedded into larger systems; they comprise heterogeneous devices combining data processing with sensing, actuating, energy scavenging, and communication and they excel in self-reliance and adaptability. What distinguishes smart systems from systems which are purely reactive is the **knowledge base**.

Gaetano Salina

Lecce 21 December 2021

The Knowledge Discovery Process (KDP)

Big Data between myth and reality

FORECASTING IN LIGHT OF BIG DATA

An extreme inductivist approach to forecasting using Big Data

According to a vaguely defined yet rather commonly held view big data may lead to dispense with theory, modeling or even hypothesizing.

All of this would be encompassed, across domains, by smart enough machine learning algorithms operating on large enough data sets.

We are interested in forecasts such that future states of a systems are predicted solely on the basis of known past states.

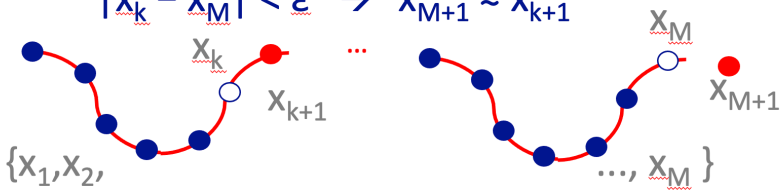
Two hypotheses are needed to give an affirmative answer:

- Similar premises lead to similar conclusions (Analogy);
- Systems which exhibit a certain behavior, will continue doing so (Determinism).

In more formal terms, given the series $\{x_1, \ldots, x_M\}$, where x_j is the vector describing the state at time $j\Delta t$, we look in the past for an analogous state, that is a vector x_k with $k < M$ "near enough" (i.e. $|x_k - x_M| < \varepsilon$, being ε the desired degree of accuracy).

Once we find such a vector, we "predict" the future at times $M + n > M$ by simply assuming for x_{M+n} the state x_{k+n}. It all seems quite easy, but it is not at all obvious that an analog can be found.

$$|x_k - x_M| < \varepsilon \rightarrow x_{M+1} \approx x_{k+1}$$



Gaetano Salina

Lecce 21 December 2021

FORECASTING IN LIGHT OF BIG DATA

arXiv:1705.11186v1 [physics.soc-ph] 31 May 2017

HYKEL HOSNI AND ANGELO VULPIANI

ABSTRACT. Predicting the future state of a system has always been a natural motivation for science and practical applications. Such a topic, beyond its obvious technical and societal relevance, is also interesting from a conceptual point of view. This owes to the fact that forecasting lends itself to two equally radical, yet opposite methodologies. A reductionist one, based on the first principles, and the naïve-inductivist one, based only on data. This latter view has recently gained some attention in response to the availability of unprecedented amounts of data and increasingly sophisticated algorithmic analytic techniques. The purpose of this note is to assess critically the role of *big data* in reshaping the key aspects of forecasting and in particular the claim that *bigger* data leads to *better* predictions. Drawing on the representative example of weather forecasts we argue that this is not generally the case. We conclude by suggesting that a clever and context-dependent compromise between modelling and quantitative analysis stands out as the best forecasting strategy, as anticipated nearly a century ago by Richardson and von Neumann.

Nothing is more practical than a good theory (L. Boltzmann)

In particular we ask whether using our knowledge of the past states of a system – and without the use of models for the evolution equation – meaningful predictions about the future are possible.

...

The key to understanding the "proper level" of abstraction lies with identifying the "relevant variables" and the effective equations which rule their time evolution.

The Knowledge Discovery Process (KDP)

Big Data between myth and reality

FORECASTING IN LIGHT OF BIG DATA

An extreme inductivist approach to forecasting using Big Data

Poincare' recurrence theorem

After a suitable time, a deterministic system with a bounded phase space returns to a state near to its initial condition.

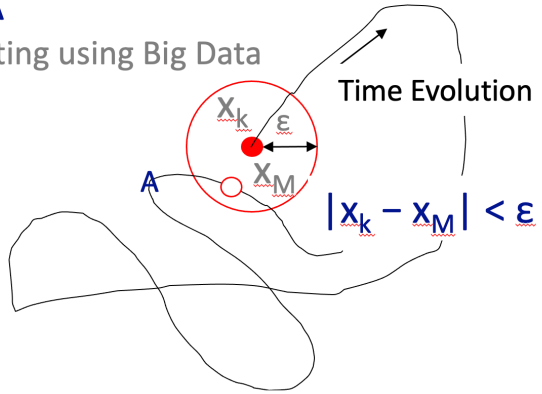
Thus an analog surely exists. How long do we have to go back to find it?

Kac who proved a Lemma to the effect that the average return time in a region A is proportional to the inverse of the probability P(A) that the system is in A.

$$\langle T_R \rangle \propto \frac{1}{P(A)}$$

Consider a system of dimension D. The probability P(A) of being in A is

$$P(A) \propto \varepsilon^D \mapsto \langle T_R \rangle = O(\varepsilon^{-D})$$



$$D \approx 10$$

$$\varepsilon \approx 0.01$$

$$\langle T_R \rangle = O(10^{20})$$

The return time is so large that in practice a recurrence is never observed.
So the required analog, whose existence is guaranteed in theory, sometimes cannot be expected to be found in practice, even if complete and precise information about the system.

Gaetano Salina

Lecce 21 December 2021

FORECASTING IN LIGHT OF BIG DATA

arXiv:1705.11186v1 [physics.soc-ph] 31 May 2017

HYKEL HOSNI AND ANGELO VULPIANI

ABSTRACT. Predicting the future state of a system has always been a natural motivation for science and practical applications. Such a topic, beyond its obvious technical and societal relevance, is also interesting from a conceptual point of view. This owes to the fact that forecasting lends itself to two equally radical, yet opposite methodologies. A reductionist one, based on the first principles, and the naïve-inductivist one, based only on data. This latter view has recently gained some attention in response to the availability of unprecedented amounts of data and increasingly sophisticated algorithmic analytic techniques. The purpose of this note is to assess critically the role of *big data* in reshaping the key aspects of forecasting and in particular the claim that *bigger* data leads to *better* predictions. Drawing on the representative example of weather forecasts we argue that this is not generally the case. We conclude by suggesting that a clever and context-dependent compromise between modelling and quantitative analysis stands out as the best forecasting strategy, as anticipated nearly a century ago by Richardson and von Neumann.

Nothing is more practical than a good theory (L. Boltzmann)

In particular we ask whether using our knowledge of the past states of a system – and without the use of models for the evolution equation – meaningful predictions about the future are possible.

...

The key to understanding the "proper level" of abstraction lies with identifying the "relevant variables" and the effective equations which rule their time evolution.

The Knowledge Discovery Process (KDP)

In socio-economical systems the gap between data and our scientific ability to actually understanding them is typically enormous. Surely the availability of huge amounts of data, sophisticated methods for its retrieval and unprecedented computational power available for its analysis will undoubtedly help moving science and technology forward.

But in spite of a persistent emphasis on a fourth paradigm (beyond experiment, theory and computation) based only on data, there is as yet no evidence data alone can bring about scientifically meaningful advance.

We therefore conclude that the big data revolution is by all means a welcome one for the new opportunities it opens.

However the role of modeling cannot be discounted.

FORECASTING IN LIGHT OF BIG DATA

arXiv:1705.11186v1 [physics.soc-ph] 31 May 2017

HYKEL HOSNI AND ANGELO VULPIANI

ABSTRACT. Predicting the future state of a system has always been a natural motivation for science and practical applications. Such a topic, beyond its obvious technical and societal relevance, is also interesting from a conceptual point of view. This owes to the fact that forecasting lends itself to two equally radical, yet opposite methodologies. A reductionist one, based on the first principles, and the naïve-inductivist one, based only on data. This latter view has recently gained some attention in response to the availability of unprecedented amounts of data and increasingly sophisticated algorithmic analytic techniques. The purpose of this note is to assess critically the role of *big data* in reshaping the key aspects of forecasting and in particular the claim that *bigger* data leads to *better* predictions. Drawing on the representative example of weather forecasts we argue that this is not generally the case. We conclude by suggesting that a clever and context-dependent compromise between modelling and quantitative analysis stands out as the best forecasting strategy, as anticipated nearly a century ago by Richardson and von Neumann.

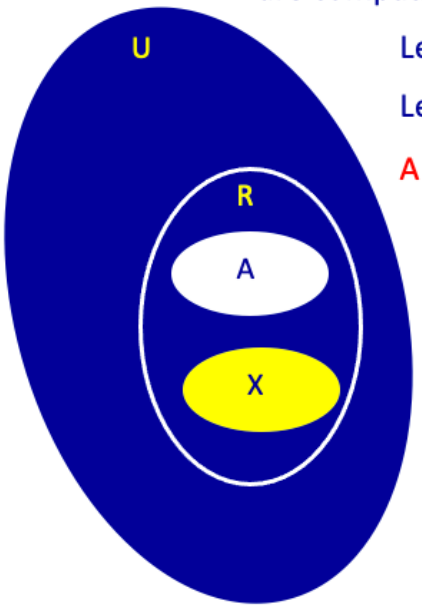
Nothing is more practical than a good theory (L. Boltzmann)

A clever and context-dependent trade-off between modeling and quantitative analysis stands out as the best strategy for meaningful prediction.
Von Neumann

The Knowledge Discovery Process (KDP)

Multiple Perceptron: Learning & Generalization

Let be U the set of all possible input-output rules, some of them are compatible with a given rule R



Let be A the set used for the learning (p examples)

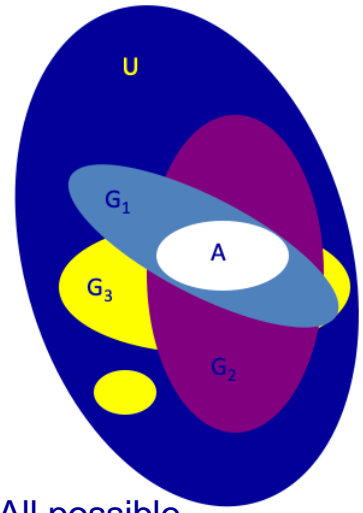
Let be X the set used for the model validation.

A e X are random choosen and are representative of R

The Network "known" A and knows nothing of X and R

Big data undoubtedly constitute a great opportunity for scientific and technological advance, with a potential for considerable socio-economic impact. To make the most of it, however, the ensuing developments at the interface of statistics, machine learning and artificial intelligence, must be coupled with adequate methodological foundations.

H. Hosni & A. Vulpiani



All possible generalizations compatible with A are valid !

When can I be sure that the dataset chosen for training allows me a unique generalization ?

It is more than a data overfitting

The Knowledge Discovery Process (KDP)

A ontological & semantic approach to heterogeneous and unstructured datasets as way to Data Definition and Acquisition.

The develop of Computational Models, i.e.

- Heuristic Models: the goal of the is matched to a particular method, such as classification, regression, or clustering the transformed data. A decision about which models and parameters might be appropriate must be done and matching a particular data mining method with the overall criteria of the KDP in Datasets process is necessary.
- Theoretical Models: the transformed data are used to derive the “interaction” parameters of a theoretical dynamic model. The KDP coincides with the understanding of the system’s physical dynamics.

as Data Mining

...

It can be, roughly speaking, the way of *physics* to try to tame complex systems

A Contribution to the Mathematical Theory of Epidemics.

By W. O. KERMACK and A. G. MCKENDRICK.

(Communicated by Sir Gilbert Walker, F.R.S.—Received May 13, 1927.)

(From the Laboratory of the Royal College of Physicians, Edinburgh.)

It was the year 1927

In this case the equations are

$$\frac{dx}{dt} = -\kappa xy$$

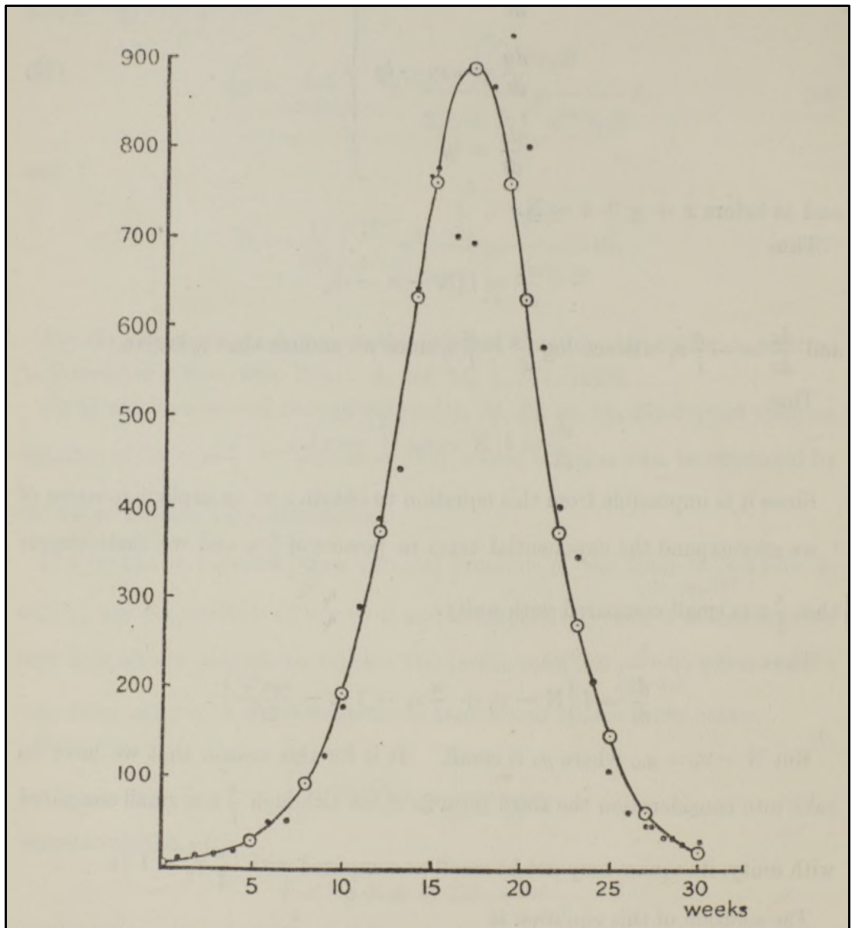
$$\frac{dy}{dt} = \kappa xy - ly$$

$$\frac{dz}{dt} = ly$$

and as before $x + y + z = N$.

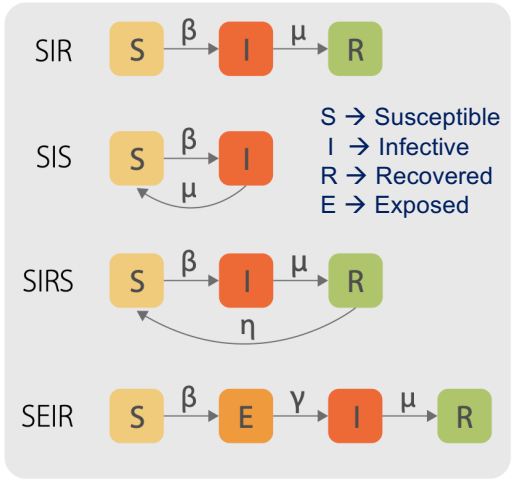
SIR Model

We are, in fact, assuming that plague in man is a reflection of plague in rats, and that with respect to the rat (1) the uninfected population was uniformly susceptible; (2) that all susceptible rats in the island had an equal chance of being infected; (3) that the infectivity, recovery, and death rates were of constant value throughout the course of sickness of each rat; (4) that all cases ended fatally or became immune; and (5) that the flea population was so large that the condition approximated to one of contact infection.



The accompanying chart is based upon figures of deaths from plague in the island of Bombay over the period December 17, 1905, to July 21, 1906. The ordinate represents the number of deaths per week, and the abscissa denotes the time in weeks. As at least 80 to 90 per cent. of the cases reported terminate fatally, the ordinate may be taken as approximately representing dz/dt as a function of t. The calculated curve is drawn from

Classic Approach: Compartmental Models



$$S + I \xrightarrow{\beta} 2I, \qquad I \xrightarrow{\mu} R.$$

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \mu I$$
$$\frac{dR}{dt} = \mu I$$

The SIR Model

Let be $\rho^I = I/\mathrm{N} \sim 0$

$$\frac{d\rho^I}{dt} \simeq (\beta - \mu)\rho^I \quad \Rightarrow \quad \rho^I(t) \simeq \rho^I(0)e^{(\beta-\mu)t}$$

The number of infectious individuals grows exponentially if

$$\beta - \mu > 0 \quad \Rightarrow \quad R_0 = \frac{\beta}{\mu} > 1$$

Reproduction Number

Epidemic processes in complex networks

Romualdo Pastor-Satorras,[1] Claudio Castellano,[2,3] Piet Van Mieghem,[4] and Alessandro Vespignani[5,6]
[1] Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain
[2] Istituto dei Sistemi Complessi (ISC-CNR), via dei Taurini 19, I-00185 Roma, Italy
[3] Dipartimento di Fisica, "Sapienza" Università di Roma, P.le A. Moro 2, I-00185 Roma, Italy
[4] Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[5] Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston MA 02115 USA
[6] Institute for Scientific Interchange Foundation, Turin 10133, Italy

(Dated: September 21, 2015)

In recent years the research community has accumulated overwhelming evidence for the emergence of complex and heterogeneous connectivity patterns in a wide range of biological and sociotechnical systems. The complex properties of real-world networks have a profound impact on the behavior of equilibrium and nonequilibrium phenomena occurring in various systems, and the study of epidemic spreading is central to our understanding of the unfolding of dynamical processes in complex networks. The theoretical analysis of epidemic spreading in heterogeneous networks requires the development of novel analytical frameworks, and it has produced results of conceptual and practical relevance. A coherent and comprehensive review of the vast research activity concerning epidemic processes is presented, detailing the successful theoretical approaches as well as making their limits and assumptions clear. Physicists, mathematicians, epidemiologists, computer, and social scientists share a common interest in studying epidemic spreading and rely on similar models for the description of the diffusion of pathogens, knowledge, and innovation. For this reason, while focusing on the main results and the paradigmatic models in infectious disease modeling, the major results concerning generalized social contagion processes are also presented. Finally, the research activity at the forefront in the study of epidemic spreading in coevolving, coupled, and time-varying networks is reported.

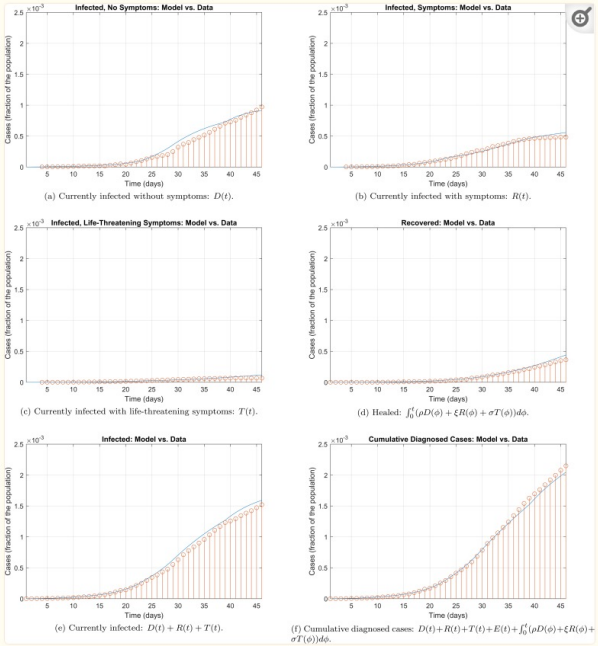
CONTENTS

https://arxiv.org/pdf/1408.2701.pdf

- An epidemic can be rephrased as a stochastic reaction-diffusion process.
- Individuals belonging to the different compartments can be represented as different kinds of "particles" or "species", that evolve according to a given set of mutual interaction rules.
- In the continuous time limit each reaction (transition) is defined by an appropriate reaction rate.

Classic Approach: Compartmental Models

We infer the model parameters based on the official data (…) about the evolution of the epidemic in Italy from 20/2/2020 through 5/4/2020.

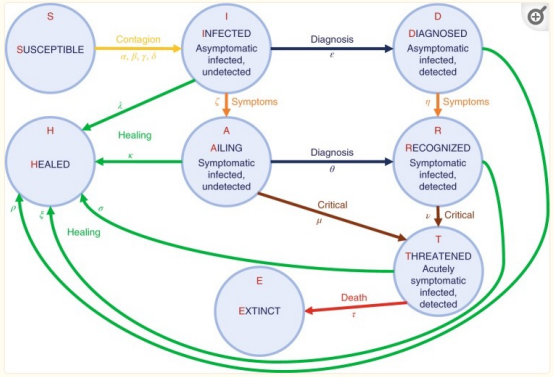


(a) Currently infected without symptoms: $D(t)$. (b) Currently infected with symptoms: $R(t)$. (c) Currently infected with life-threatening symptoms: $T(t)$. (d) Healed: $\int_0^t (\rho D(\phi) + \xi R(\phi) + \sigma T(\phi))d\phi$. (e) Currently infected: $D(t) + R(t) + T(t)$. (f) Cumulative diagnosed cases: $D(t)+R(t)+T(t)+E(t)+\int_0^t (\rho D(\phi)+\xi R(\phi)+\sigma T(\phi))d\phi$.

Extended Data Fig. 3

Model simulation compared to real data.

Comparison between the official data (red dots histogram) and the results with the calibrated SIDARTHE model (blue line). Panel (**a**): number of reported infected with no (or mild) symptoms, who are quarantined at home. Panel (**b**): number of reported infected with symptoms, who are hospitalised. Panel (**c**): number of reported infected with life-threatening symptoms, admitted to ICU. Panel (**d**): number of reported recovered individuals. Panel (**e**): total number of reported infected in all categories. Panel (**f**): number of cumulative reported cases.



SIDARTHE Model

Fig. 1

The model.

Graphical scheme representing the interactions among different stages of infection in the mathematical model SIDARTHE: S, susceptible (uninfected); I, infected (asymptomatic or pauci-symptomatic infected, undetected); D, diagnosed (asymptomatic infected, detected); A, ailing (symptomatic infected, undetected); R, recognized (symptomatic infected, detected); T, threatened (infected with life-threatening symptoms, detected); H, healed (recovered); E, extinct (dead).

$$
\begin{aligned}
\dot{S}(t) &= -S(t)(\alpha I(t) + \beta D(t) + \gamma A(t) + \delta R(t)) \\
\dot{I}(t) &= S(t)(\alpha I(t) + \beta D(t) + \gamma A(t) + \delta R(t)) - (\epsilon + \zeta + \lambda) I(t) \\
\dot{D}(t) &= \epsilon I(t) - (\eta + \rho) D(t) \\
\dot{A}(t) &= \zeta I(t) - (\theta + \mu + \kappa) A(t) \\
\dot{R}(t) &= \eta D(t) + \theta A(t) - (\nu + \xi) R(t) \\
\dot{T}(t) &= \mu A(t) + \nu R(t) - (\sigma + \tau) T(t) \\
\dot{H}(t) &= \lambda I(t) + \rho D(t) + \kappa A(t) + \xi R(t) + \sigma T(t) \\
\dot{E}(t) &= \tau T(t)
\end{aligned}
$$

Nat Med. 2020 Apr 22 : 1–6.
doi: 10.1038/s41591-020-0883-7 [Epub ahead of print]
PMCID: PMC7175834
PMID: 32322102

Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy

Giulia Giordano,[✉1] Franco Blanchini,[2] Raffaele Bruno,[3,4] Patrizio Colaneri,[5,6] Alessandro Di Filippo,[3] Angela Di Matteo,[3] and Marta Colaneri[3]

Associated Data

Supplementary Materials

Data Availability Statement

Abstract Go to:

In Italy, 128,948 confirmed cases and 15,887 deaths of people who tested positive for SARS-CoV-2 were registered as of 5 April 2020. Ending the global SARS-CoV-2 pandemic requires implementation of multiple population-wide strategies, including social distancing, testing and contact tracing. We propose a new model that predicts the course of the epidemic to help plan an effective control strategy. The model considers eight stages of infection: susceptible (S), infected (I), diagnosed (D), ailing (A), recognized (R), threatened (T), healed (H) and extinct (E), collectively termed SIDARTHE. Our SIDARTHE model discriminates between infected individuals depending on whether they have been diagnosed and on the severity of their symptoms. The distinction between diagnosed and non-diagnosed individuals is important because the former are typically isolated and hence less likely to spread the infection. This delineation also helps to explain misperceptions of the case fatality rate and of the epidemic spread. We compare simulation results with real data on the COVID-19 epidemic in Italy, and we model possible scenarios of implementation of countermeasures. Our results demonstrate that restrictive social-distancing measures will need to be combined with widespread testing and contact tracing to end the ongoing COVID-19 pandemic.

Subject terms: Epidemiology, Infectious diseases, Computational models, Differential equations

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175834/

A New Approach: Complex Networks Models

The Complex Networks Zoo

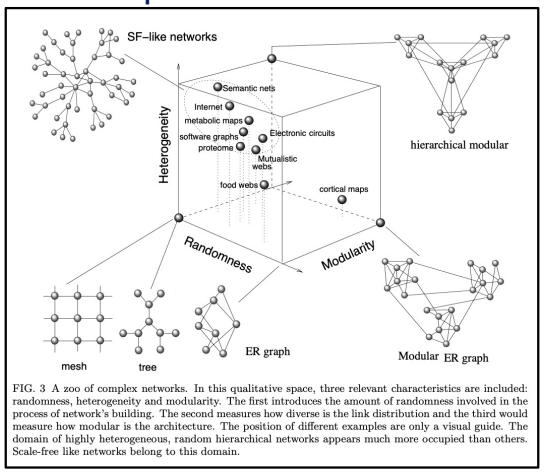


FIG. 3 A zoo of complex networks. In this qualitative space, three relevant characteristics are included: randomness, heterogeneity and modularity. The first introduces the amount of randomness involved in the process of network's building. The second measures how diverse is the link distribution and the third would measure how modular is the architecture. The position of different examples are only a visual guide. The domain of highly heterogeneous, random hierarchical networks appears much more occupied than others. Scale-free like networks belong to this domain.

https://www.researchgate.net/publication/225782040_Information_Theory_of_Complex_Networks_On_Evolution_and_Architectural_Constraints/figures?lo=1

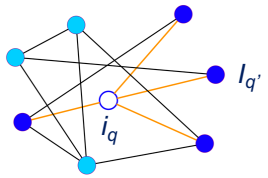
SIR: $\mu=1$ and $\lambda=\beta$. Use two equations

$$\frac{dr_q(t)}{dt} = i_q(t)$$

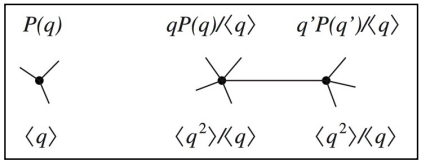
Probability that a given link connects to susceptible nodes.

$$\frac{di_q(t)}{dt} = -i_q(t) + \lambda q\left[1-i_q(t)\right]\sum_{q'}\frac{q'-1}{q'}P(q'|q)i_{q'}(t)$$

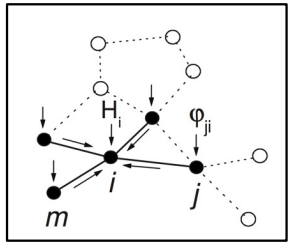
The contribution of the first neighbours nodes



(q'-1)/q' ratio is due to the fact that an infected vertex in this model cannot infect back its infector, and so one of the *q* edges is effectively blocked.

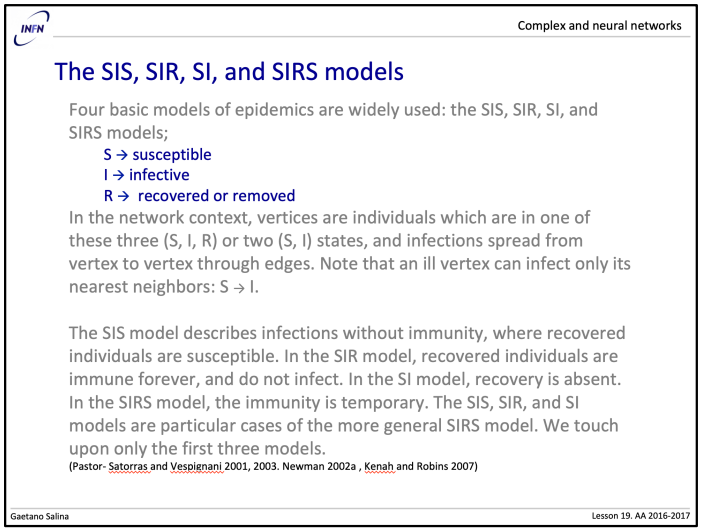


Distribution of connections and the mean degree of a randomly chosen vertex (left) differ sharply from those of end vertices of a randomly chosen edge (right)



$$P(q'|q) = \frac{q'P(q')}{\langle q\rangle}$$

The network is completely described by the joint degree-degree distribution *P(q',q)* and by *N* . It is convenient to use a conditional probability *P(q'|q)* that if an end vertex of an edge has degree *q*, then the second end has degree *q'*.



Complex and neural networks

The SIS, SIR, SI, and SIRS models

Four basic models of epidemics are widely used: the SIS, SIR, SI, and SIRS models;

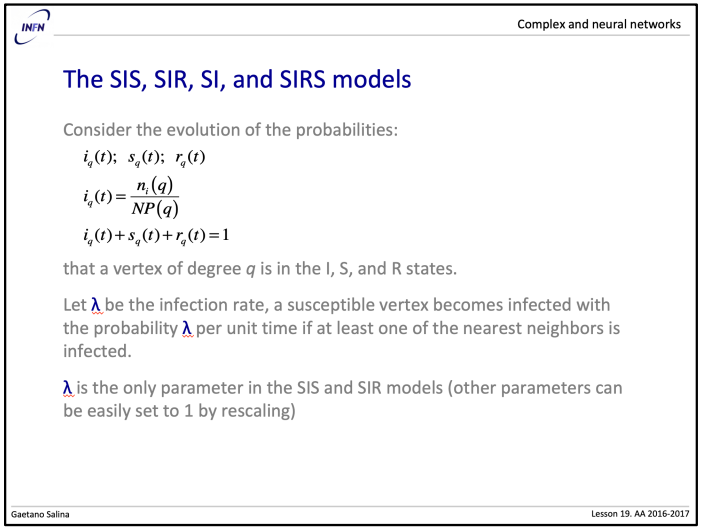
- S → susceptible
- I → infective
- R → recovered or removed

In the network context, vertices are individuals which are in one of these three (S, I, R) or two (S, I) states, and infections spread from vertex to vertex through edges. Note that an ill vertex can infect only its nearest neighbors: S → I.

The SIS model describes infections without immunity, where recovered individuals are susceptible. In the SIR model, recovered individuals are immune forever, and do not infect. In the SI model, recovery is absent. In the SIRS model, the immunity is temporary. The SIS, SIR, and SI models are particular cases of the more general SIRS model. We touch upon only the first three models.

(Pastor- Satorras and Vespignani 2001, 2003. Newman 2002a , Kenah and Robins 2007)

Gaetano Salina — Lesson 19. AA 2016-2017



Complex and neural networks

The SIS, SIR, SI, and SIRS models

Consider the evolution of the probabilities:

$$i_q(t);\ s_q(t);\ r_q(t)$$

$$i_q(t) = \frac{n_i(q)}{NP(q)}$$

$$i_q(t) + s_q(t) + r_q(t) = 1$$

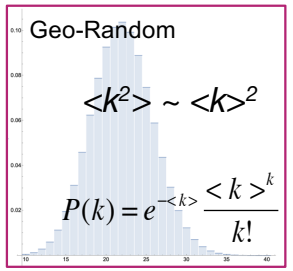
that a vertex of degree *q* is in the I, S, and R states.

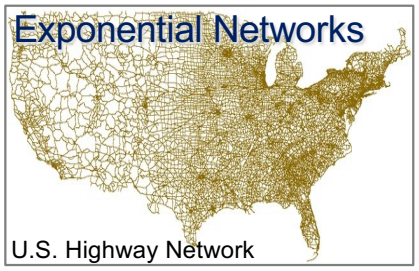
Let λ be the infection rate, a susceptible vertex becomes infected with the probability λ per unit time if at least one of the nearest neighbors is infected.

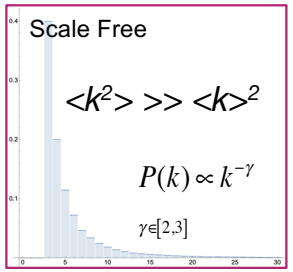
λ is the only parameter in the SIS and SIR models (other parameters can be easily set to 1 by rescaling)

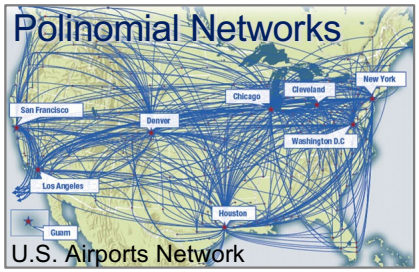
Gaetano Salina — Lesson 19. AA 2016-2017

A New Approach: Complex Networks Models









Epidemic thresholds:

The epidemic threshold λ_c is a basic notion in epidemiology, the fractions of infected and recovered or removed vertices in the final state (stationary) are defined as:

$$i(\infty) = \sum_q P(q) i_q(t \to \infty)$$
$$r(\infty) = \sum_q P(q) r_q(t \to \infty)$$

$\partial_t i(\infty) = 0$ In the final, stationary, state we have (SIS):

$$\sum_q P(q) i_q(t) = \sum_q P(q) \left\{ -i_q(t) + \lambda q \left[1 - i_q(t)\right] \sum_{q'} \frac{q' P(q')}{< q >} i_{q'}(t) \right\} =$$

$$-i(t) + \lambda i(t)\left[1 - i(t)\right] \left\{ \sum_{q,q'} \frac{q' P(q') q P(q)}{< q >} \right\} = -i(t) + \lambda i(t)\left[1 - i(t)\right] \frac{< q^2 >}{< q >} = 0$$

Geo-Random $< q^2 > \approx < q >^2$

$$\partial_t i(t) = 0 \rightarrow \begin{cases} i = 0 & \lambda < \lambda_c \\ i \approx \lambda - \lambda_c & \lambda \geq \lambda_c \end{cases}$$

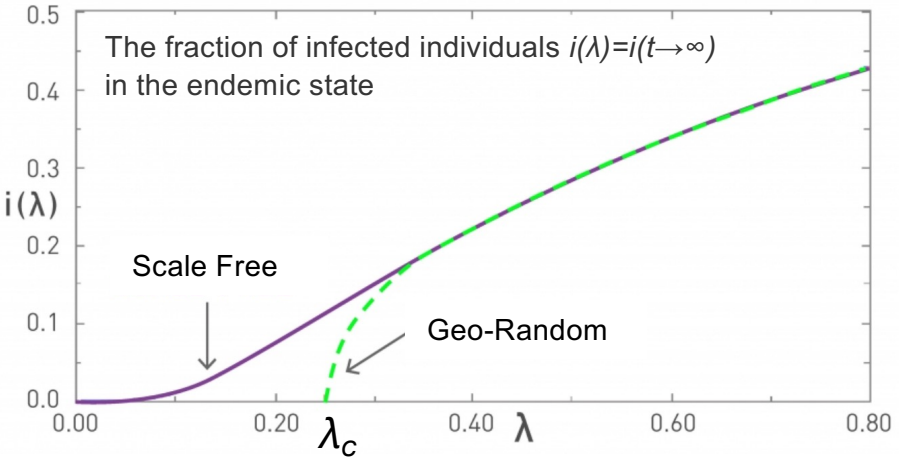
$\lambda_c = <q>^{-1}$

Reproduction Number $R_0 = \langle k \rangle \frac{\beta}{\mu}$

Scale Free $< q^2 > \;>>\; < q >$

$$\partial_t i(t) = 0 \rightarrow i \approx \lambda$$

No epidemic threshold



The Geo-Random has a finite epidemic threshold λ_c. If the spreading rate λ exceeds λ_c, the pathogen becomes endemic.
For a Scale-Free network we have λ_c=0, hence even viruses with a very small spreading rate λ can persist in the population.

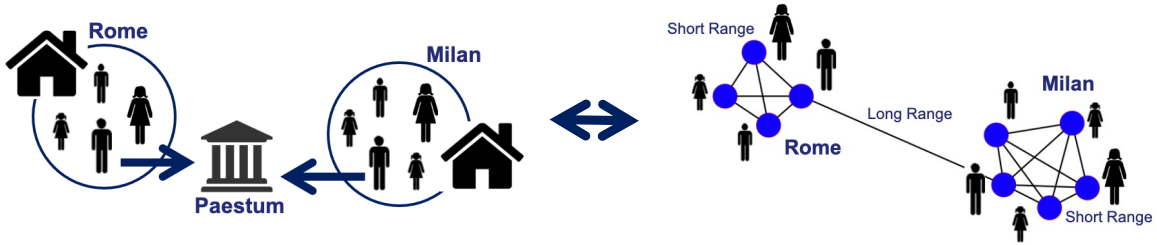
The SIS models on Uncorrelated Network

A New Approach: Complex Networks Models

Go beyond the time evolution:

Metric Spaces and Network Topology

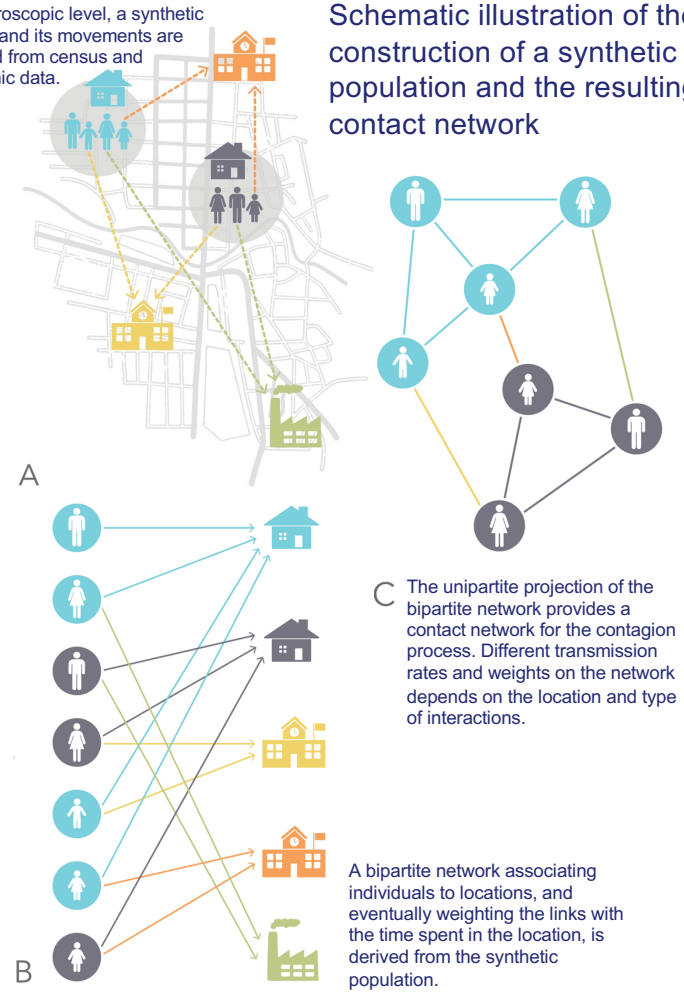
- The Compartmental Models assume random and homogeneous mixing, where each member in any compartment is similar and indistinguishable from the others.
- In reality, each individual has his own social contact network over which diseases propagate, usually differing from that of other members in a compartment.
- Individuals move in a metric space. Once a spatial scale has been set, the social contact network can be divided into short or long interactions.



- Diekmann et al. (2012) show the weakness of R_0 by studying a line and square lattice topology and they conclude that network and percolation theory needs to be consulted to compute the epidemic threshold, leading to a new definition of the basic reproduction number depending on the topology / metric of the network.

At the macroscopic level, a synthetic population and its movements are constructed from census and demographic data.

Schematic illustration of the construction of a synthetic population and the resulting contact network



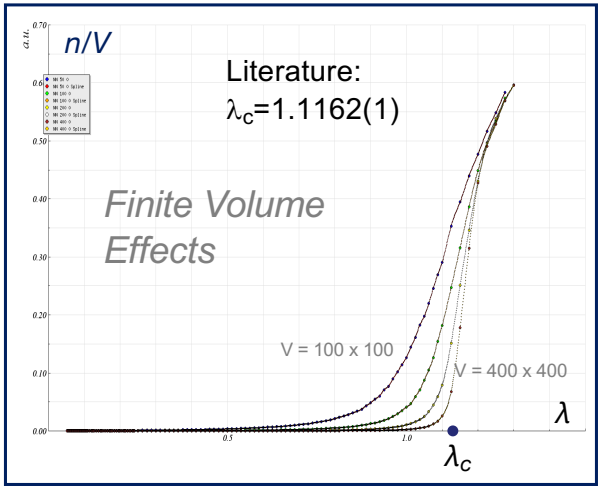
The unipartite projection of the bipartite network provides a contact network for the contagion process. Different transmission rates and weights on the network depends on the location and type of interactions.

A bipartite network associating individuals to locations, and eventually weighting the links with the time spent in the location, is derived from the synthetic population.

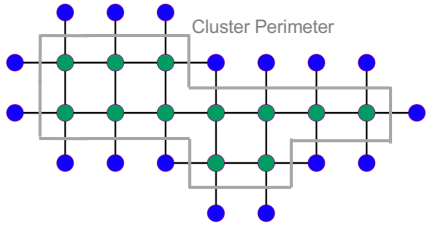
https://arxiv.org/pdf/1408.2701.pdf

Go beyond the time evolution:

Metric Spaces and Network Topology

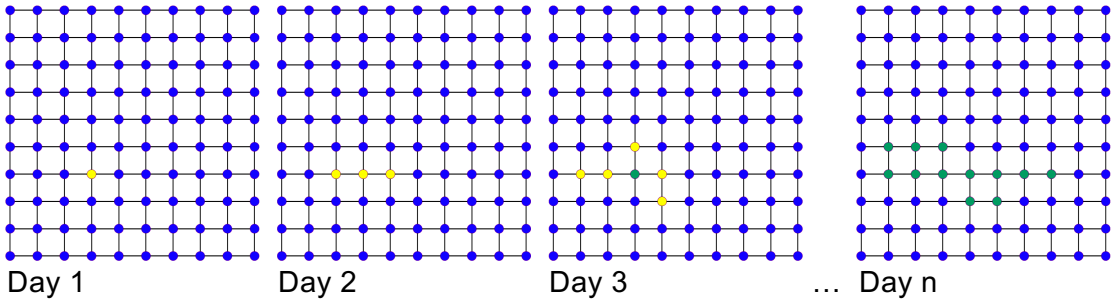


S
I
R

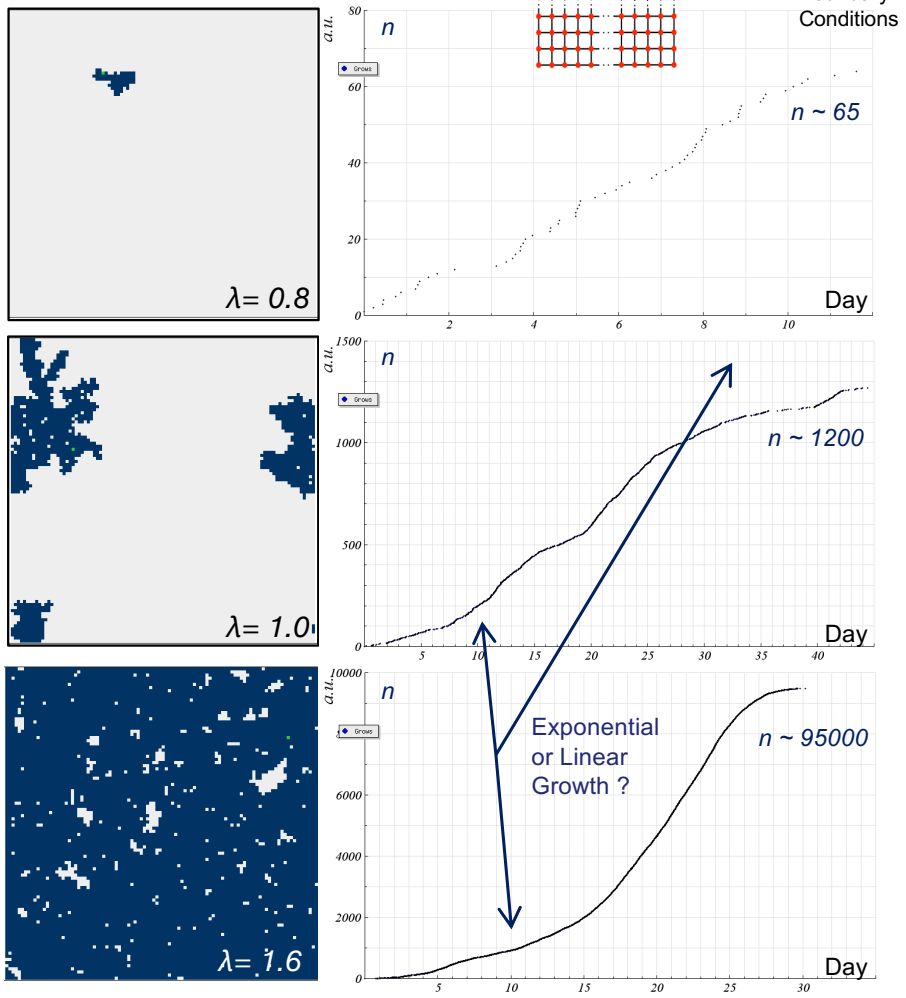


Cluster Volume:
n = Number of (● || ●)

Cluster Perimeter:
m = Number of (●—● || ●—●)



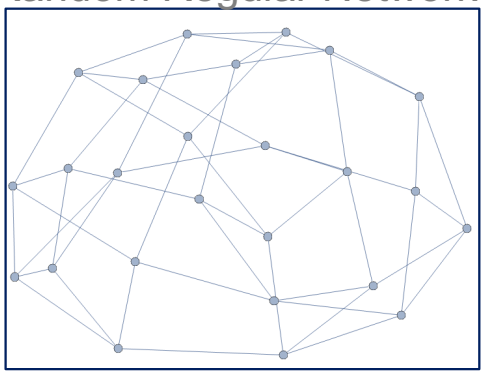
SIR models on 2D Grid



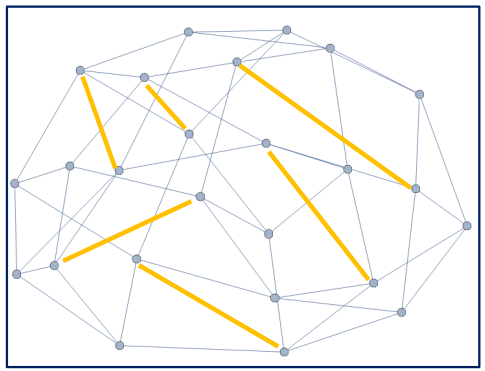
Go beyond the time evolution:

Metric Spaces and Network Topology

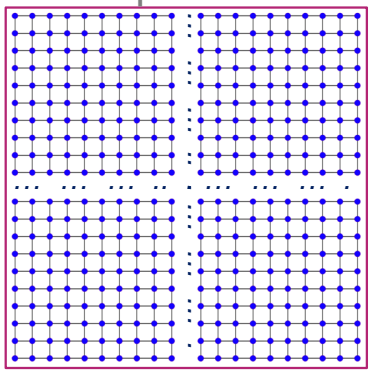
ER Random Regular Network



ER Network

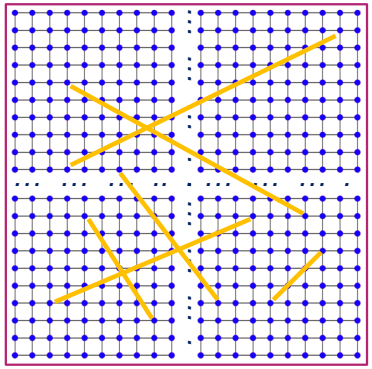


2D Square Lattice

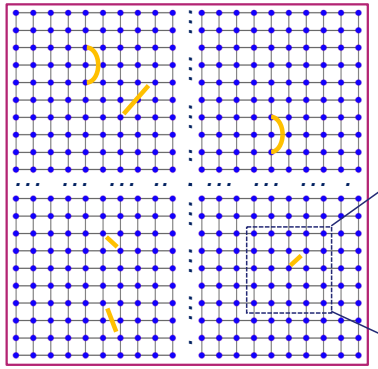


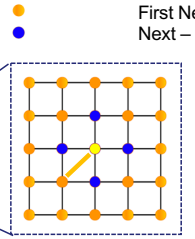
Closed Boundary Conditions

2D SW Network



2D NN Network

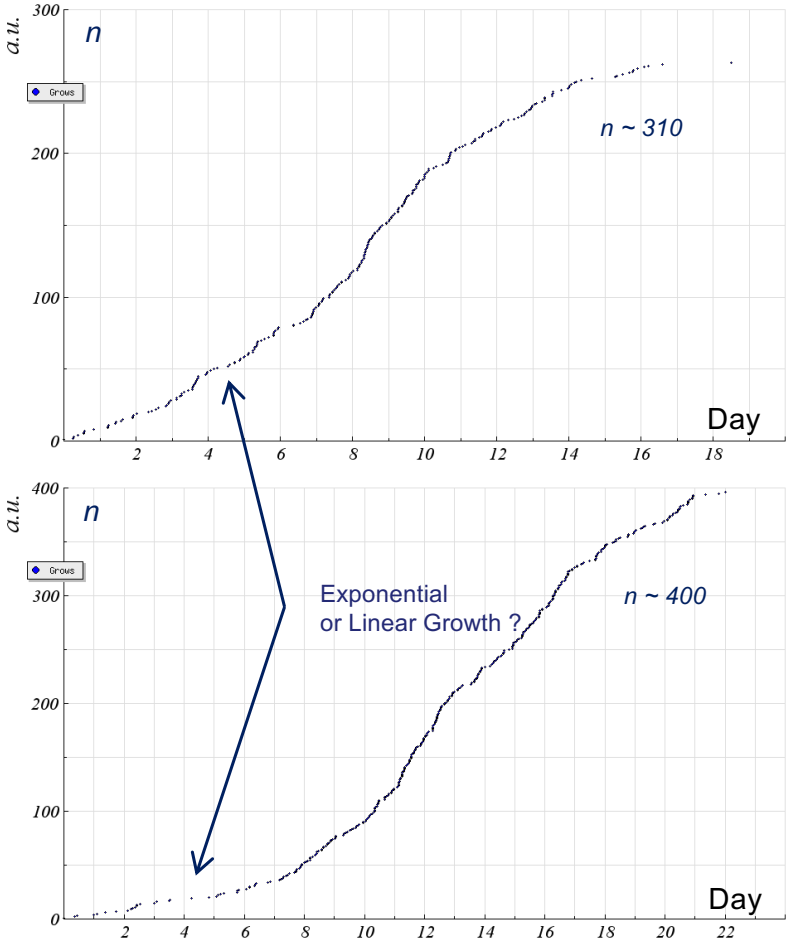


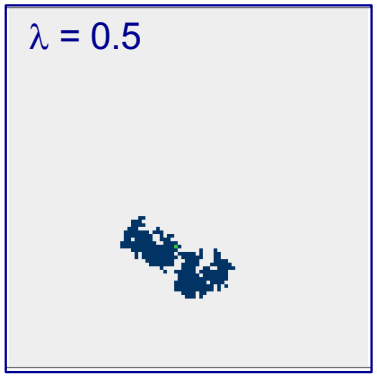


SIR models on 2D Networks

- Start from a 2D Square Lattice and a Erdos-Renyi (ER) Regular Random Graph. $\langle q \rangle_{SL} = \langle q \rangle_{ER} = 4$; $\langle d \rangle_{SL} = \langle d \rangle_{ER} = \langle q \rangle_{SL} / 2$.
- Adding the same links number to defined graphs, we obtain:
  - Erdos-Renyi (ER) Random Graph.
  - 2D Small Word (SW) Network
  - 2D Next-Next Neighbours (NN) Network
- $\langle d \rangle_{ER} = \langle d \rangle_{ER} = \langle d \rangle_{ER}$

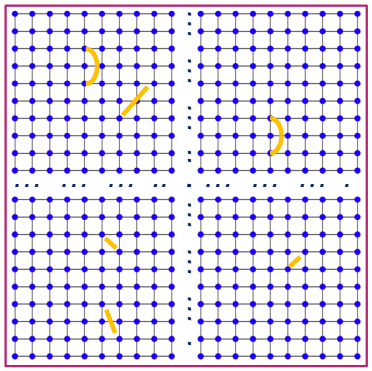
SIR models on 2D Networks



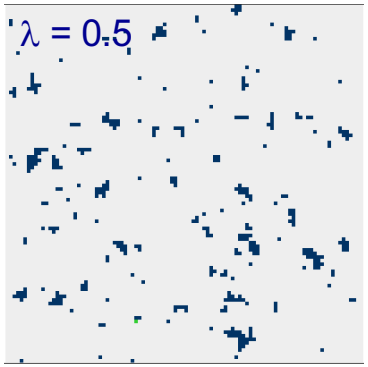


Next-Next Links; Size = 100; λ_c = 0.5; +Links: 30 %

2D NN Network

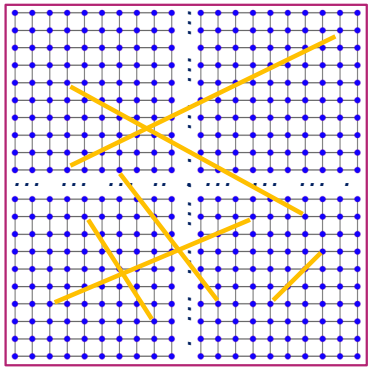


$< d >_{SW} = < d >_{NN}$



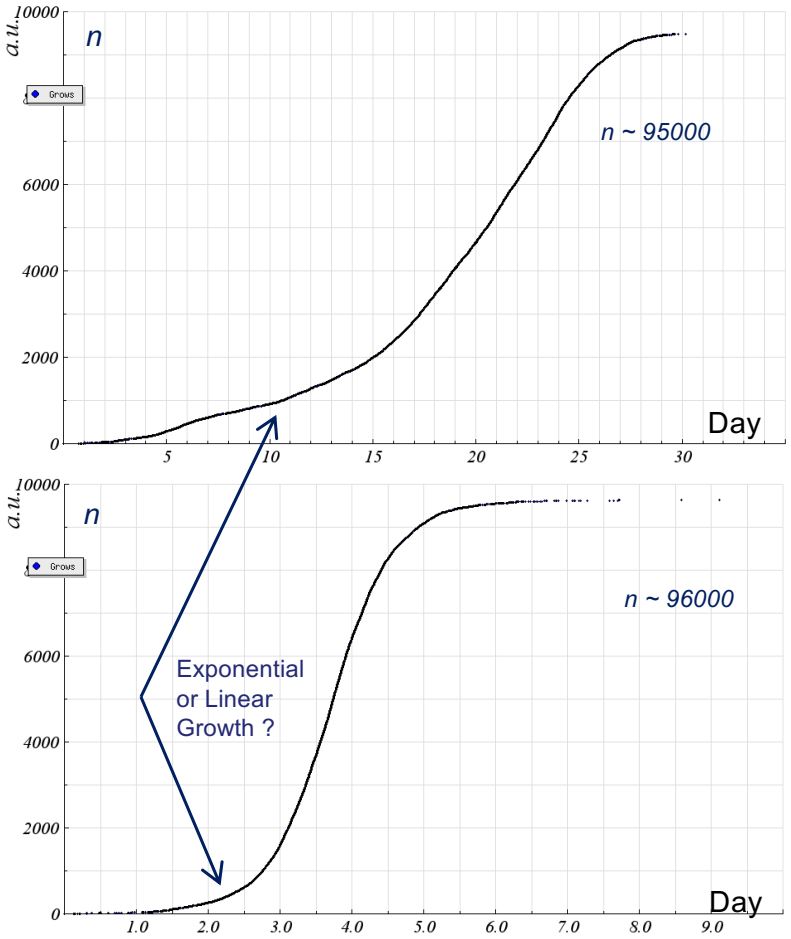
Small World Links; Size = 100; λ = 0.5; +Links: 30 %

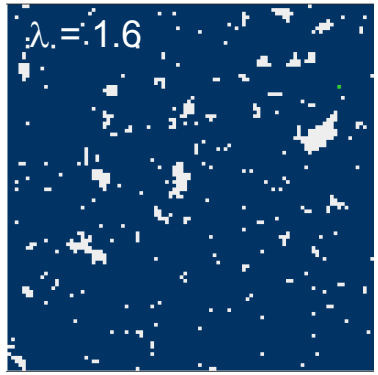
2D SW Network



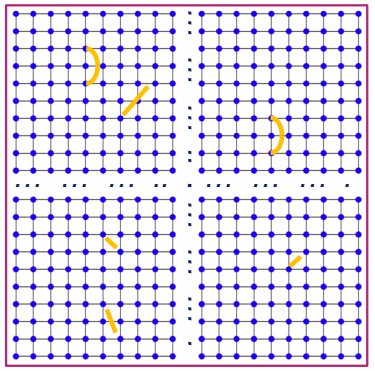
Different spread in Space

SIR models on 2D Networks





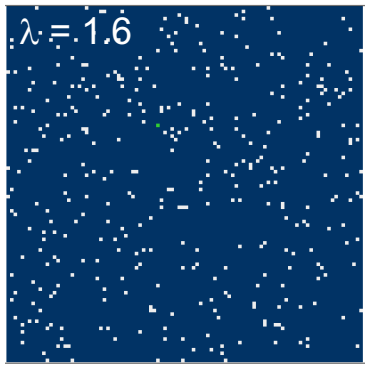
2D NN Network

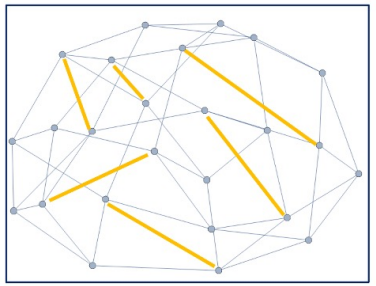


Next-Next Links; Size = 100; λ_c = 1.6; +Links: 30 %

$< d >_{SW} = < d >_{NN}$

ER Network





ER Links; Size = 100; λ = 1.6; +Links: 30 %

Different spread in Time

SIR models on 2D Networks

+Links = 0 %

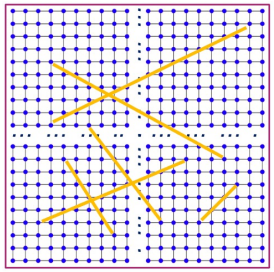
$$_{ER}\lambda_c \neq {}_{NN}\lambda_c = {}_{SW}\lambda_c$$

+Links = 6 %

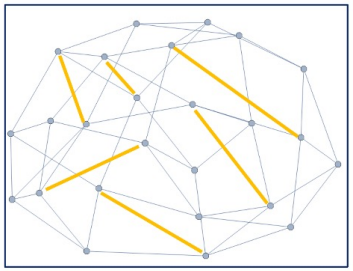
$$_{ER}\lambda_c \neq {}_{NN}\lambda_c \neq {}_{SW}\lambda_c$$

+Links = 50 %

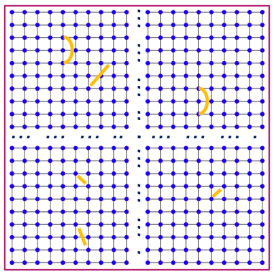
$$_{ER}\lambda_c \neq {}_{NN}\lambda_c = {}_{SW}\lambda_c$$



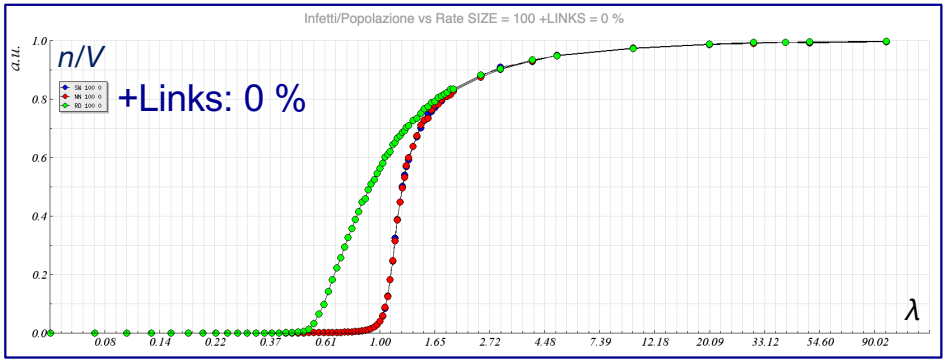
2D SW Network

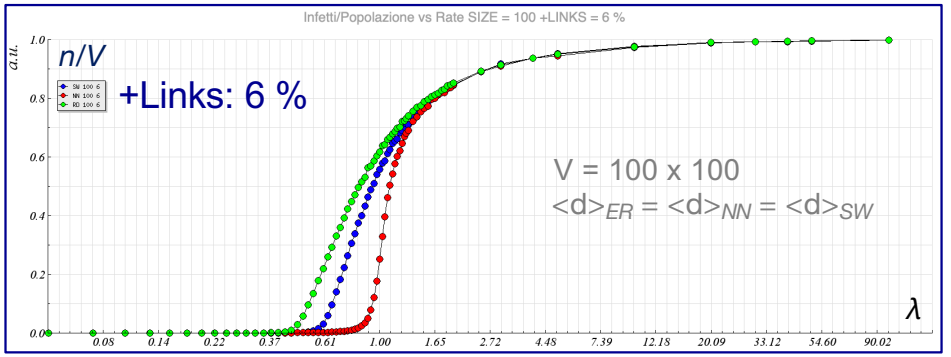


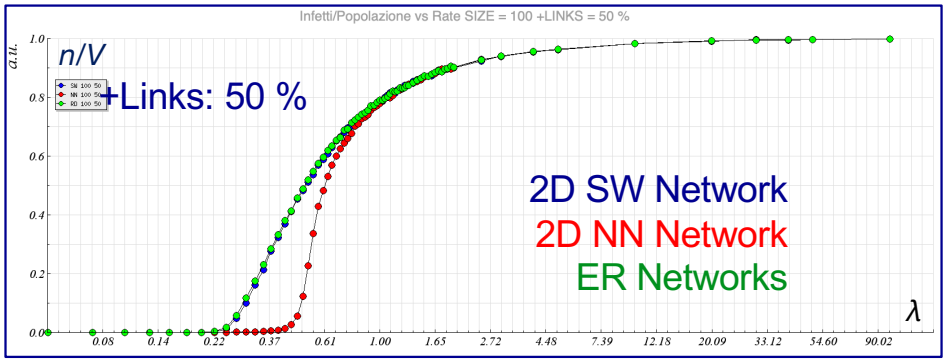
ER Network



2D NN Network



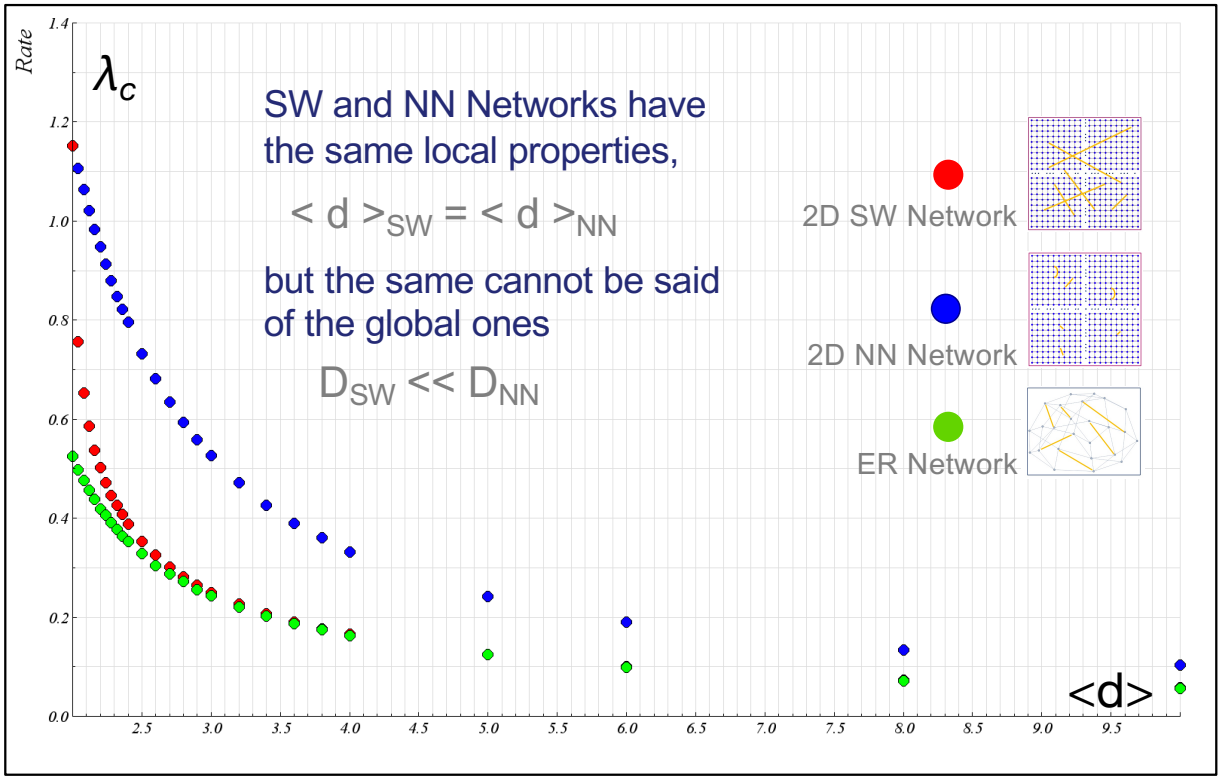




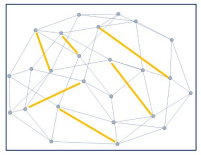
SW Networks quickly forget the 2D Square Lattice substrate.

In the limit of infinite links added, SW Networks behaves like ER Networks.

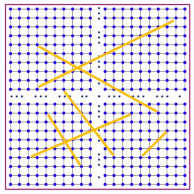
SIR models on 2D Networks



A graph's diameter D is the largest number of vertices which must be traversed in order to travel from one vertex to another when paths which backtrack, detour, or loop are excluded from consideration.



The probability to have k-loops goes to zero in thermodynamic limit, i.e. *p*(k-loops) → 0. Statistically irrelevant giant loops are present.

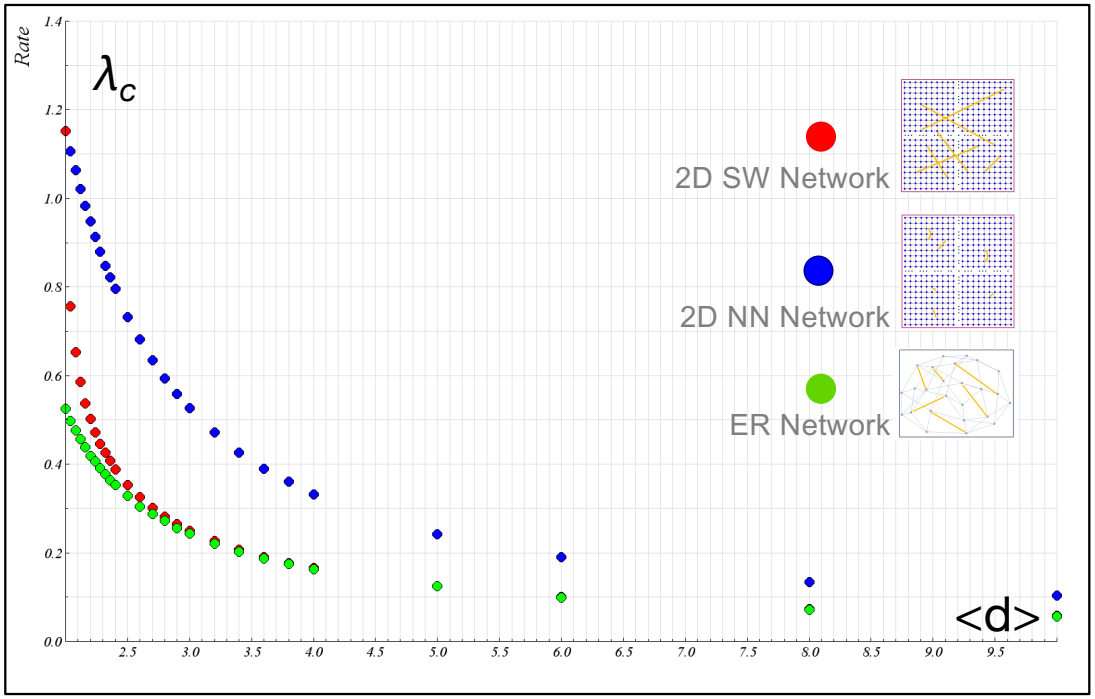


Adding SW links, the open m-paths grows, while the number of loops remains constant (V 4-loops). Due to SW links, statistically irrelevant giant loops are present.

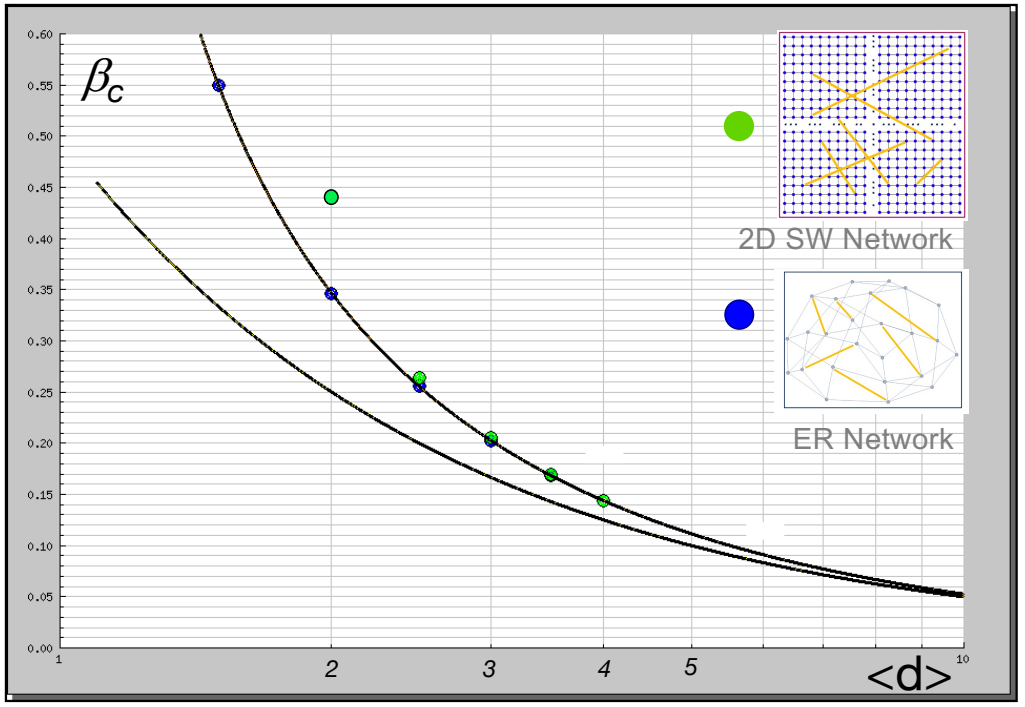
In the limit of infinite links added, SW Networks behaves as ER Networks.

SW Networks quickly forget the 2D Square Lattice substrate.

SIR models on 2D Networks



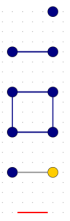
Ising Model on 2D Networks



In different physical models we observe the same effects due to the topological and metric properties of the network underlying the dynamics of the models.

Go beyond the time evolution:

Metric Spaces and Network Topology



	A	B
n	18	18
k	17	18
l	0	1
m	38	36
p	38	36
χ	1	1

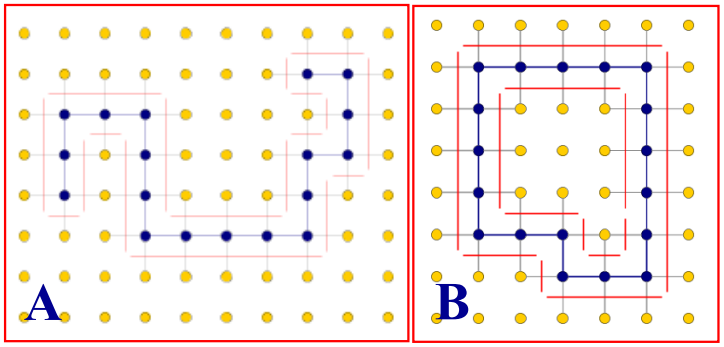
$m = p$

$$\chi = n - k + l$$

$$m = 2dn - 2k$$

$$m = 2n(d-1) - 2l + 2\chi$$

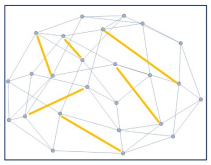
$\chi =$ **Euler Characteristic**



SIR models on 2D Networks

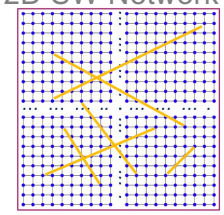
- Open and Closed Paths
- Clusters Perimeter versus Cluster Volume

ER Network



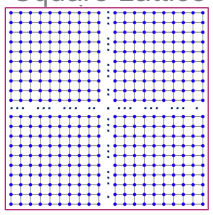
The probability to have k-loops goes to zero in thermodynamic limit, i.e. p(k-loops) → 0. Statistically irrelevant giant loops are present. ER Networks can be seen as Trees.

2D SW Network



Adding SW links, the open m-paths grows, while the number of loops remains constant (we have V 4-loops). Due to SW links, statistically irrelevant giant loops are present. SW Networks can be seen as Trees.

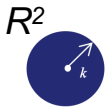
2D Square Lattice



In thermodynamic limit the ratio k-loops over k-paths remains constant, so that the effects of closed loops cannot be neglected.

Go beyond the time evolution:

Metric Spaces and Network Topology

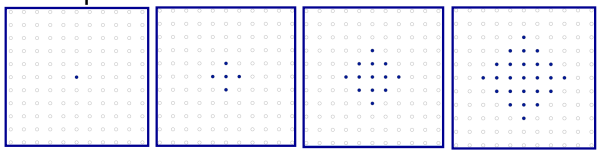


$n = \pi k^2$

$m = \pi k$

$m = (\pi n)^{1/2}$

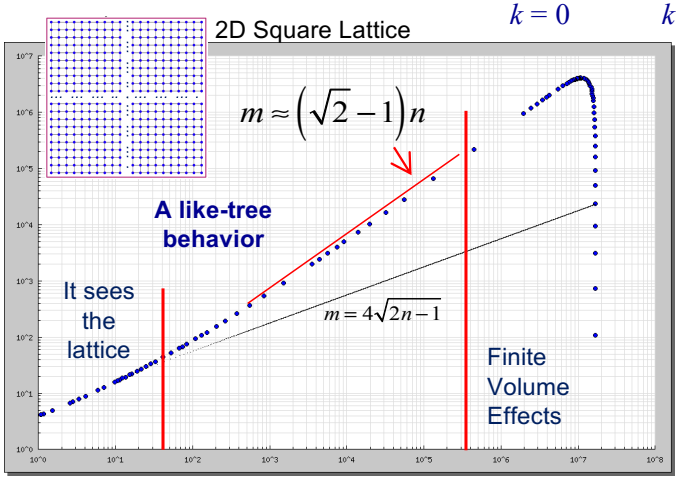
2D Square Lattice

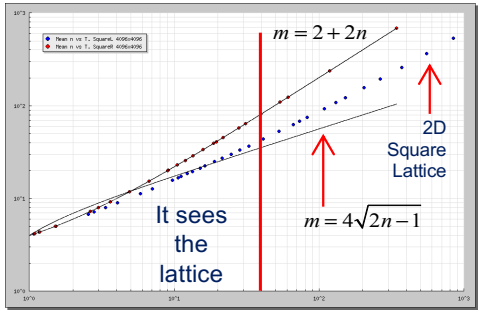


$$n = 1 + 2k^2 + 2k$$

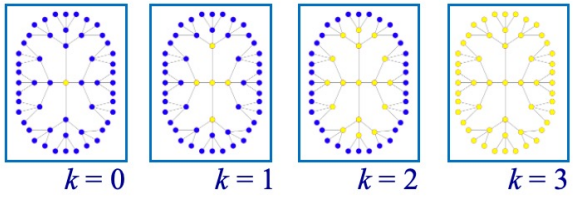
$$m = 8k + 4$$

$$m = 4\sqrt{2n-1}$$





ER Network



$$n = 2\left(3^k\right) - 1$$

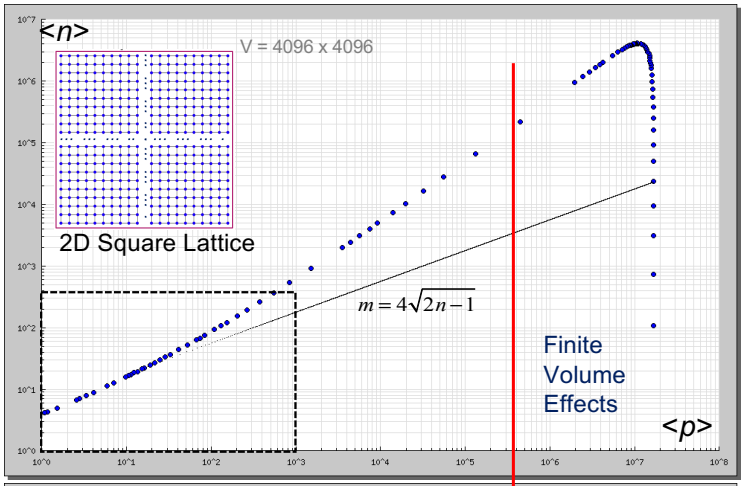
$$m = 4\left(3\right)^k$$

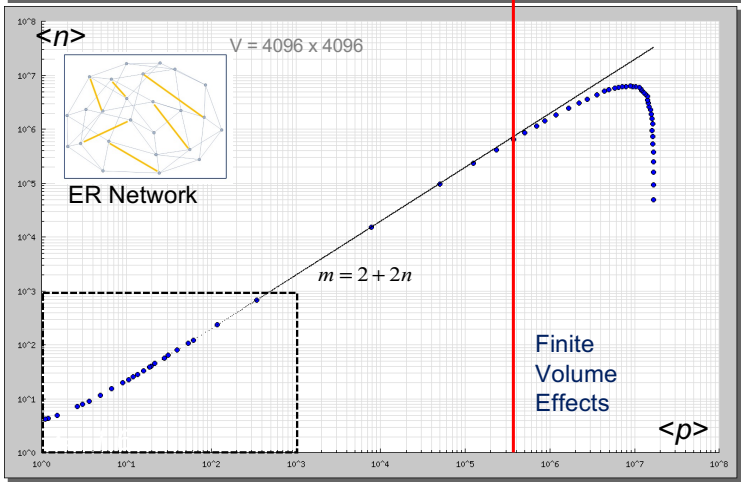
$$m = 2 + 2n$$

SIR models on 2D Networks

Clusters Perimeter versus Cluster Volume

Closed Boundary Conditions





Go beyond the time evolution:

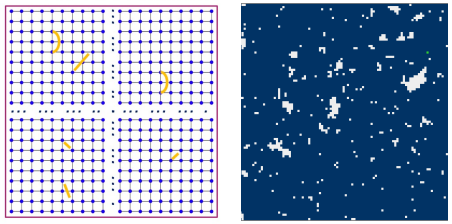
Metric Spaces and Network Topology

The impact of heterogeneous connectivity patterns, reflected by an underlying network topology, on epidemic behavior is a key point and must be understood.

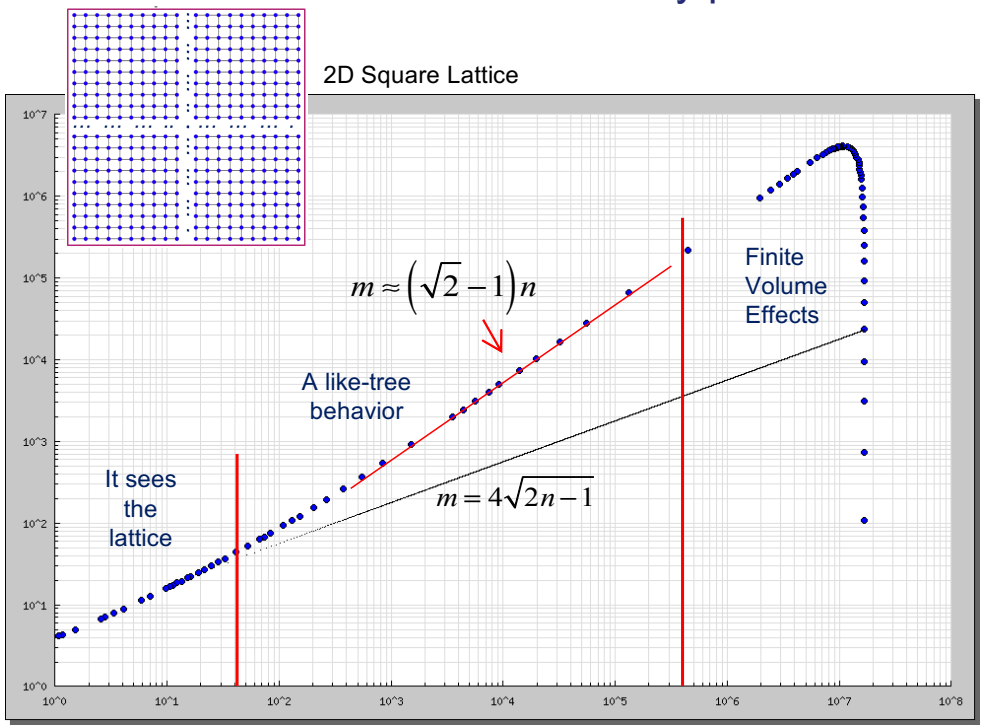
SIR models on 2D Networks

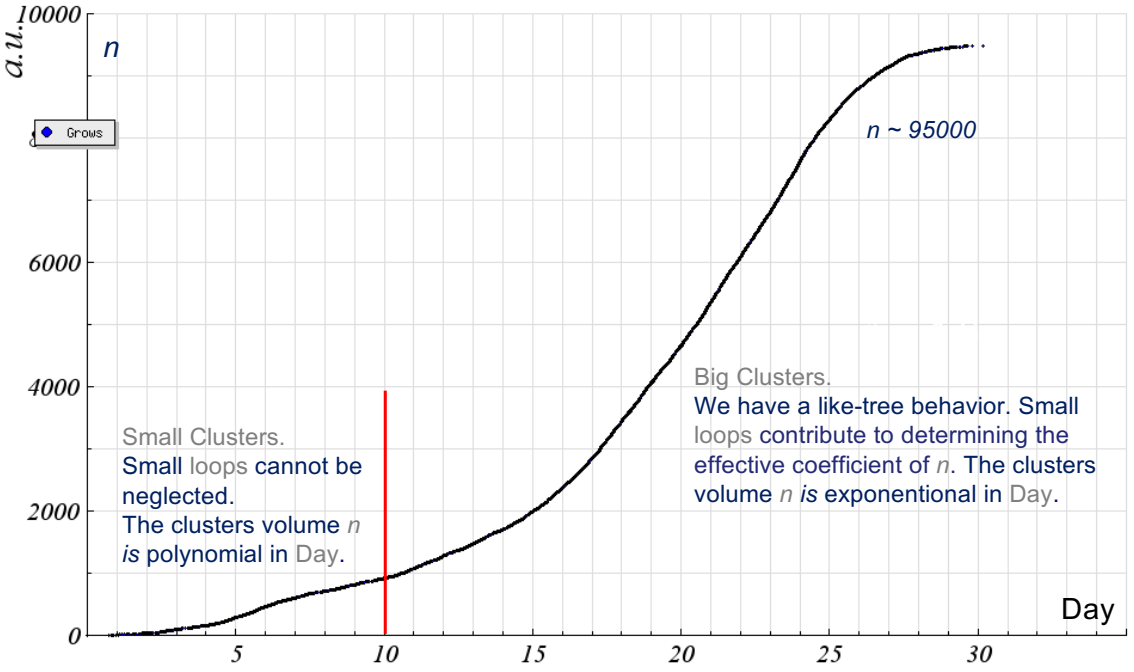
Clusters Perimeter versus Cluster Volume

2D NN Network



Size = 100;
λ = 1.6;
+Links: 30 %

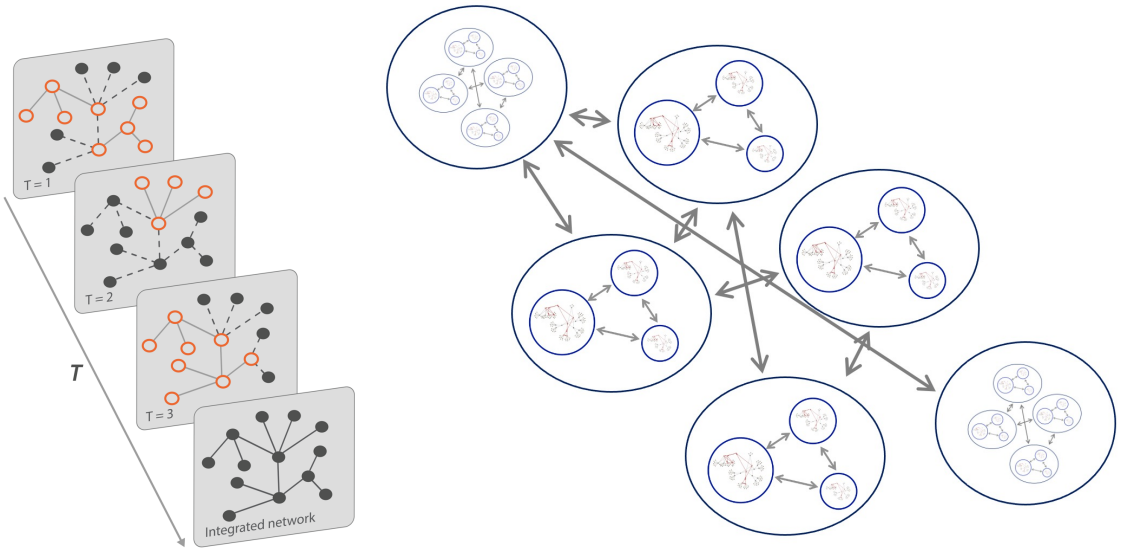




Go beyond the time evolution:

Metric Spaces and Network Topology

Time-dependent hierarchy structure, with both temporal and spatial scales to be determined.

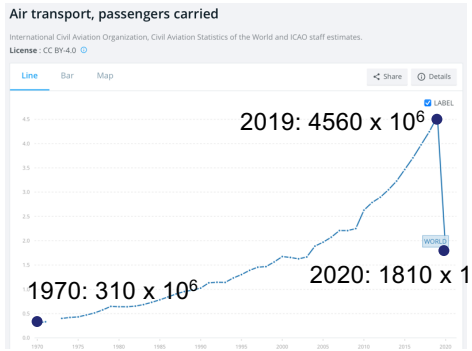


Modeling human mobility is madatory for the construction of a realistic theory of the space-time spread of an epidemic.

The Hong Kong flu was a flu pandemic whose outbreak in 1968 and 1969 killed between one and four million people globally.

The first recorded instance of the outbreak appeared on 13 July 1968 in British Hong Kong.

The first recorded case of the outbreak appeared in Italy was in January 1969, six months later.



The complexity and heterogeneity of the present time human mobility network favor considerably the global spreading of infectious diseases. Only unfeasible mobility restrictions reducing the global travel fluxes by 90% or more would be effective.

Epidemic on complex networks and definition of a Knowledge Base for the analysis of the Covid19 in Italy

G. Salina
INFN Sezione di Roma Tor Vergata

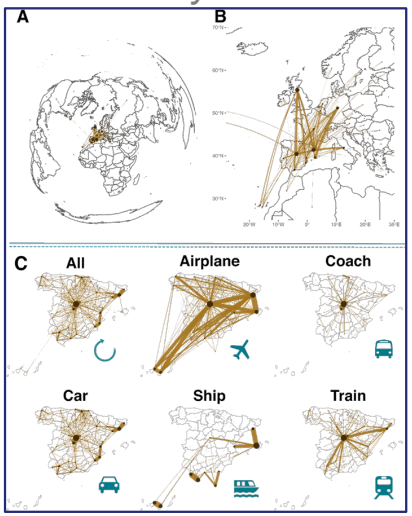
G. Cimini, A. Desiderio and L. Rossi Mori
Physics Department, Tor Vergata University, Rome

A. Gabrielli
Engineering Department, Roma Tre University, Rome

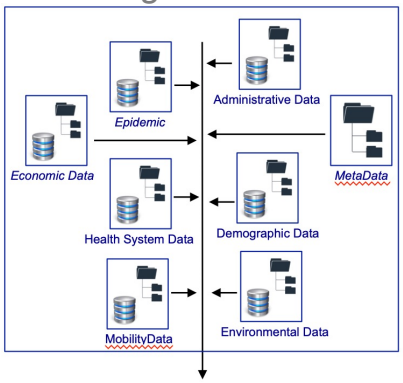
F. Sylos Labini
Enrico Fermi Research Center, CREF

L. Vigliano
Mathematics Department, Tor Vergata University, Rome

Networks Dynamics



Knowledge Base



Networks Dynamics

Multiplex mobility network and metapopulation epidemic simulations of Italy based on Open Data

The patterns of human mobility play a key role in the spreading of infectious diseases and thus represent a key ingredient of epidemic modeling and forecasting.

Unfortunately, as the Covid-19 pandemic has dramatically highlighted, for the vast majority of countries there is no availability of granular mobility data (not only).

We build a multiplex mobility network based solely on open data, and implement a SIR metapopulation model that allows scenario analysis through data-driven stochastic simulations.

The mobility flows that we estimate are in agreement with real-time proprietary data from smartphones (as Facebook Data for Good).

Our modeling approach can thus be useful in contexts where high-resolution mobility data is not available.

Multiplex mobility network and metapopulation epidemic simulations of Italy based on Open Data

Antonio Desiderio
Physics Department, University of Rome "Tor Vergata", 00133 Rome (Italy) and Centro Ricerche Enrico Fermi, 00184 Rome (Italy)

Giulio Cimini
Physics Department, University of Rome "Tor Vergata", 00133 Rome (Italy) Centro Ricerche Enrico Fermi, 00184 Rome (Italy) and Istituto Nazionale di Fisica Nucleare, Sezione Roma "Tor Vergata", 00133 Rome (Italy)

Gaetano Salina
Istituto Nazionale di Fisica Nucleare, Sezione Roma "Tor Vergata", 00133 Rome (Italy)

(Dated: May 10, 2022)

The patterns of human mobility play a key role in the spreading of infectious diseases and thus represent a key ingredient of epidemic modeling and forecasting. Unfortunately, as the Covid-19 pandemic has dramatically highlighted, for the vast majority of countries there is no availability of granular mobility data. This hinders the possibility of developing computational frameworks to monitor the evolution of the disease and to adopt timely and adequate prevention policies. Here we show how this problem can be addressed in the case study of Italy. We build a multiplex mobility network based solely on open data, and implement a SIR metapopulation model that allows scenario analysis through data-driven stochastic simulations. The mobility flows that we estimate are in agreement with real-time proprietary data from smartphones. Our modeling approach can thus be useful in contexts where high-resolution mobility data is not available.

I. INTRODUCTION

Facing the SARS-CoV-2 outbreak has required an unprecedented effort for mankind. Governments around the world have applied containment measures in multiple occasions to hinder the spread of the virus. In particular, as human interactions and mobility are known to deeply affect the patterns of epidemic spreading [1–3], non-pharmaceutical interventions (NPIs) have been widely adopted to reduce contacts between individuals [4–6].

NPIs can be generally divided in two main categories [7]. Bottom-up NPIs are self-initiated measurements and behavioral changes due to increase risk perception of the populations [8–10], which can be measured via surveys or mobile data [11–13]. Top-down NPIs are measurements imposed by governments whose effectiveness can be measured via changes in the mobility [14–16] and reduction of the disease prevalence [17, 18]. Making informed decisions, both at the top-down and bottom-up levels, requires timely and quality data about the current stage of the outbreak and contact patterns of individuals [19, 20]. In particular, these data serve to inform computational epidemic models [21], which produce detailed scenario analyses that can be fundamental to inform strategies of response and mitigation of a disease [3, 22–27] and have been widely exploited to face Covid-19 [28–41].

Epidemic models vary in complexity [21], ranging from graph-based [42], agent-based [43] and structured metapopulation models [44], each with different data requirements. A common ingredient though is the complex network modeling of social contacts, through which contagion can spread at different scales [45–47]. Mobility data are crucial in this respect [48]. Unfortunately, in many cases the health authorities had to deal with the scarcity and incompleteness of this type of data, which especially in the early stages of the pandemic prevented them from understanding the spreading patterns of the disease and the effects of the various containment measures implemented.

Here we are interested in the case study of Italy, which was the first outbreak of Covid-19 among Western countries and consequently has been the subject of many research studies that tried to model and predict the evolution of the epidemics. In addition to the development of improved compartment models calibrated on the characteristics of Covid-19 [49–55], much attention has been devoted to the cross-analysis of epidemic and mobility data, in order to evaluate the direct and indirect effect as well as the effectiveness of NPIs.

For instance, [56] developed a mean-field approach that allows to relate the amount of daily cases to the average number of people that an individual meets per day. The empirical studies have instead used mobile phone data from either Facebook Data For Good [57–59] or telecom operators [60–62]. These data are available as aggregates over territorial units, therefore they have been used to study mobility at the geographic level of regions or provinces. In particular, the authors of [58] have considered the Italian regions as separate entities, due to low level of commuting between regions, while arguing that such inter-regional mobility had a crucial role in the long-range propagation of epidemics [63]. In addition, these proprietary data represent only proxies of the mobility

https://arxiv.org/pdf/2205.03639.pdf

Networks Dynamics

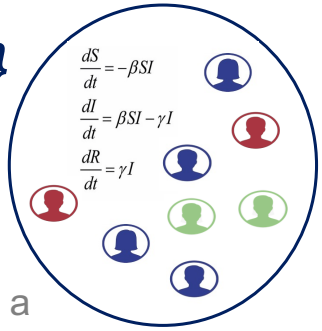
Mobility Network

The mobility network is defined by representing each of the M = 7897 municipalities as a node. We use ISTAT data to collect information on identification number, reference province, population, surface, latitude and longitude of each municipality.
https://www.istat.it/it/archivio/6789

Modelling the Municipalities

$M_i = \{ i, n_i, \rho_i, Pr_i \ldots \}$

i = 1,2,3, …,7897



The evolution of the disease is described by a SIR Metapopulation Model

Modelling the mobility of individuals

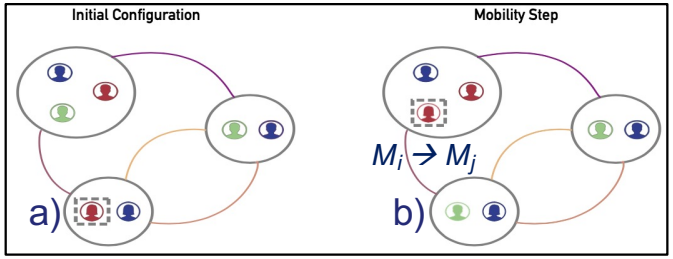
At each time step of the dynamics we:

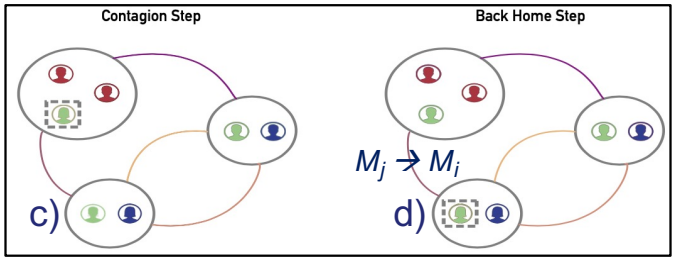
a) A sample a fraction p_M = 0.1 of the population of each municipality is choosen

b) Move it to other municipalities. This moving population is distributed over the four layers with the p_{Layer}

c) Execute a Metapopulation Dynamic step

d) At the end of the current time step, the moving population returns to the municipality of departure.





Networks Dynamics

Mobility Network

The connectivity of the Intra-Province and Inter-Province layers is set by the territorial contiguity matrix collected by ISTAT.

https://www.istat.it/it/ archivio/157423

$c_{i,j} = \begin{cases} = 1 & \text{Adjacent Municipalities} \\ = 0 & \text{Not Adjacent Municipalities} \end{cases}$

i,j = 1,2,3, …,7897

1) Intra-Province Layer

All $M_i \leftrightarrow M_j$ *if* $c_{i,j} = 0$ *and* $Pr_i = Pr_j$

2) Inter-Province Layer

All $M_i \leftrightarrow M_j$ *if* $c_{i,j} = 0$ *and* $Pr_i \neq Pr_j$

Modelling the short range networks:

Intra-province and Inter-province layers

Connections in these layers are weighted using gravity-like assumptions:

$$w_{ij}^{\mathrm{L}} = \frac{\rho_i \rho_j}{r_{ij}} c_{ij}^{\mathrm{L}}$$

where ρ_i is the population density of node i, $r_{i,j}$ is the distance between the two municipalities and $\mathrm{L} \in \{\mathrm{Intra}, \mathrm{Inter}\}$ denotes the layer.

In order to transform such weights into Markovian connection probabilities we use the following normalization

$$\tilde{w}_{ij}^{\mathrm{L}} = \frac{w_{ij}^{\mathrm{L}}}{\sum_{l \in n_i^{\mathrm{L}}} w_{il}^{\mathrm{L}}} c_{ij}^{\mathrm{L}}$$

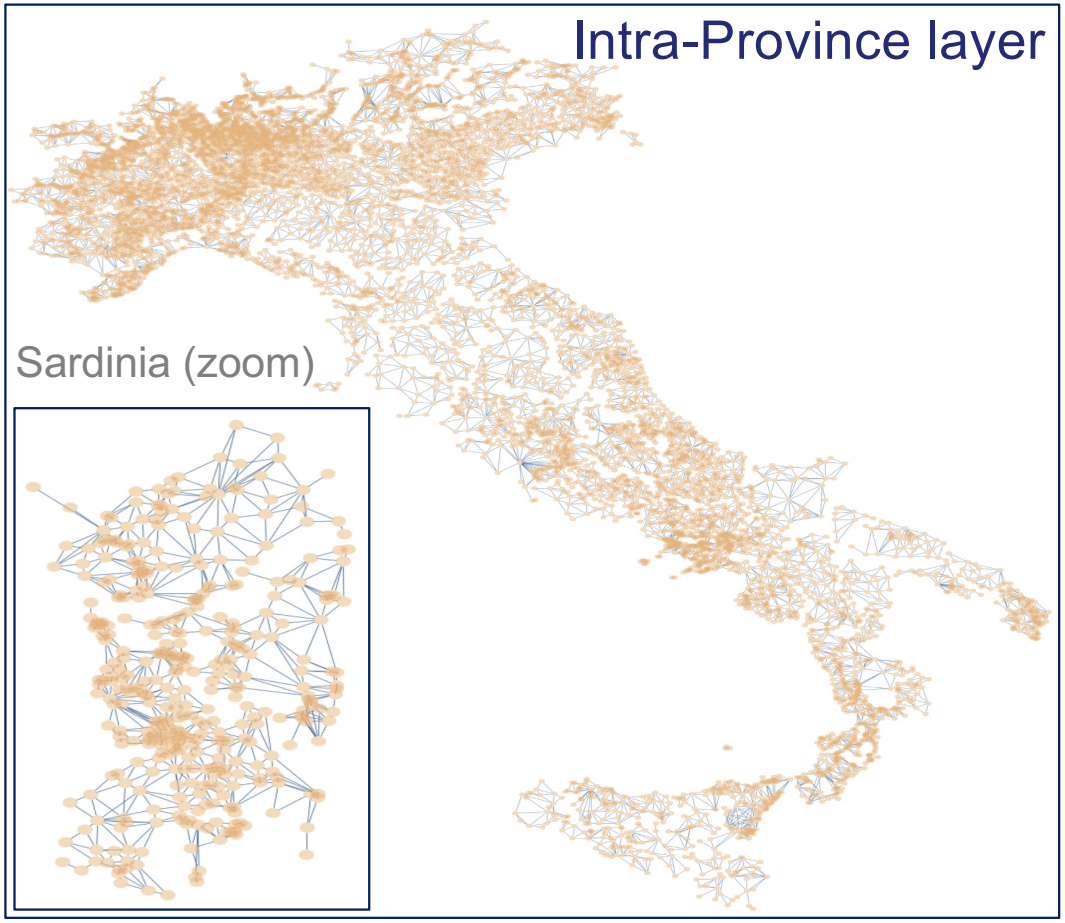
where n^L_i are the territorial neighbours of node i in the corresponding layer.

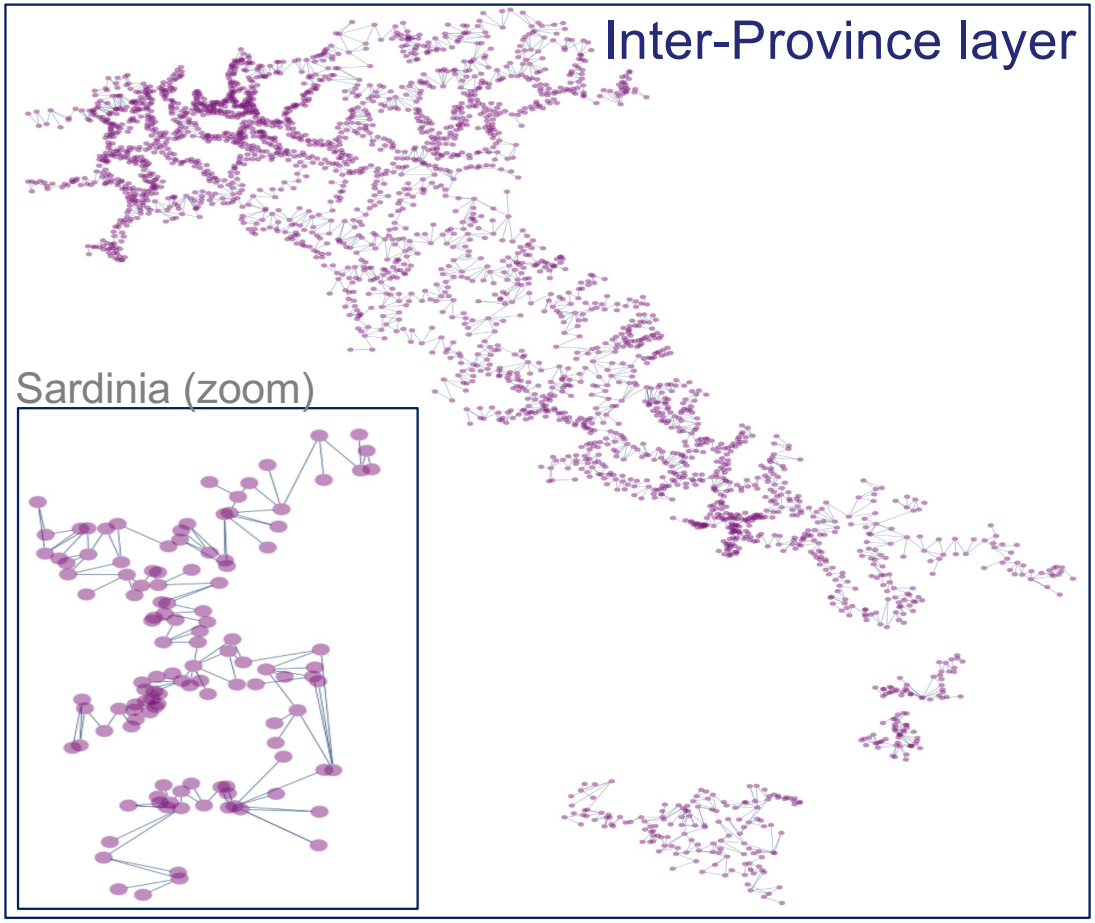
Networks Dynamics

Mobility Network

Modelling the short range networks:

Intra-province and Inter-province layers





Networks Dynamics

Mobility Network

Modelling the long range networks:

Train layer

The Trains layer

We use the RESTful API of Viaggiatreno to retrieve information.

http://www.viaggiatreno.it/

- Each station is identified by a id, latitude and longitude
- Each train departing from station i we have the id j of the next stop station and the type of connection *e* (Frecciarossa, Intercity, Metropolitana, EuroNight, Regionale, EuroCity)
- We associated each station with the municipalities that are located no more than 5 km away (in this way, on average a single station corresponds to three municipalities).

$$w_{ij}^{\mathrm{Train}} = \sum_e \tau_{ij}^e g^e$$

$$\tilde{w}_{ij}^{\mathrm{L}} = \frac{w_{ij}^{\mathrm{L}}}{\sum_{l \in n_i^{\mathrm{L}}} w_{il}^{\mathrm{L}}} \tau_{ij}^e$$

L = Train

τ_{ij}^e is boolean indicator for the presence of a connection

g^e is the fraction of total travellers for train type *e*. Dati ISTAT

https://www.istat.it/it/archivio/13995

Networks Dynamics

Mobility Network

Modelling the long range networks:
Flights layer

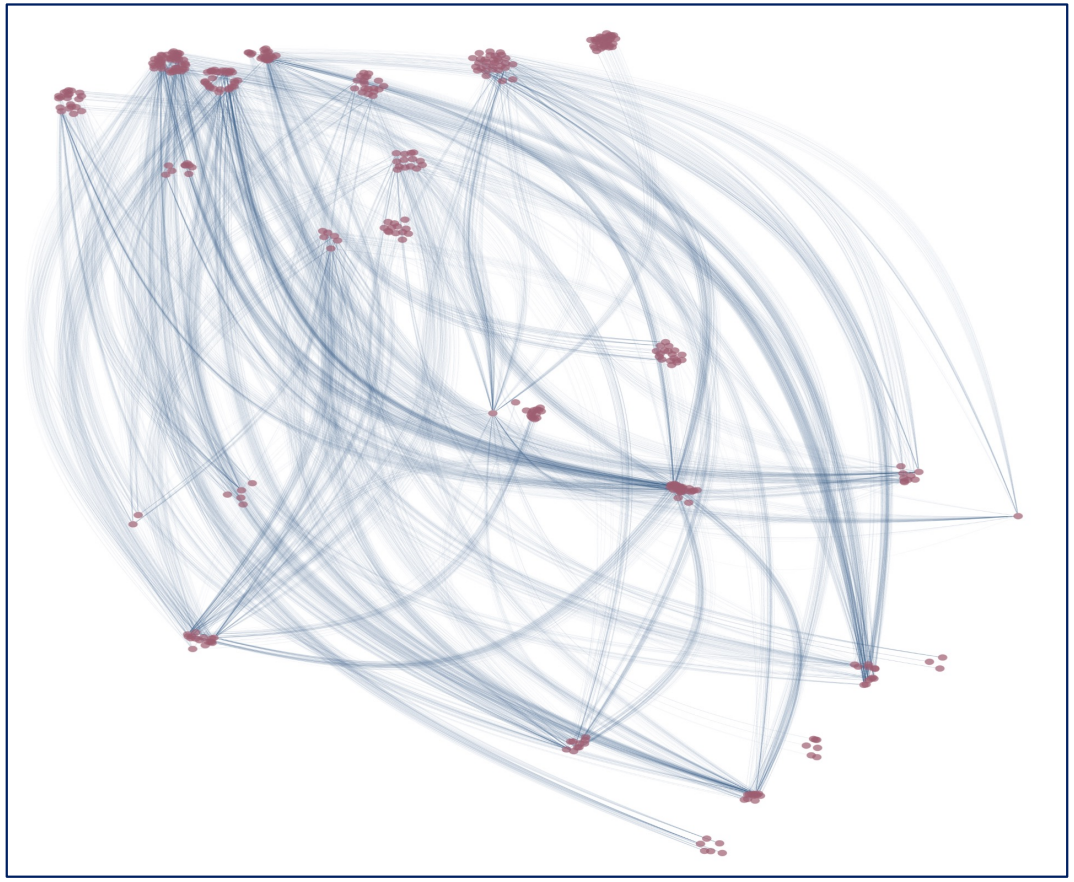
The Flights layer

- To build the Flight layer, we collect IATA and ICAO codes, latitudes and longitudes of airports from Open- Flights
  https://openflights.org/data.html
- From the annual ENAC 2019 report we collected the air traffic data: origin and destination airports, with IATA and ICAO Ids, and total passengers moving between airports.
  https://www.enac.gov.it/pubblicazioni/dati-di-traffico-2019
- We associated each airport with all municipalities located no more than 15 km away (on average a single airport is mapped on ten municipalities).

$$w_{ij}^{\text{Flight}} = g_{ij} \qquad \tilde{w}_{ij}^{\text{L}} = \frac{w_{ij}^{\text{L}}}{\sum_{l \in n_i^{\text{L}}} w_{il}^{\text{L}}}$$

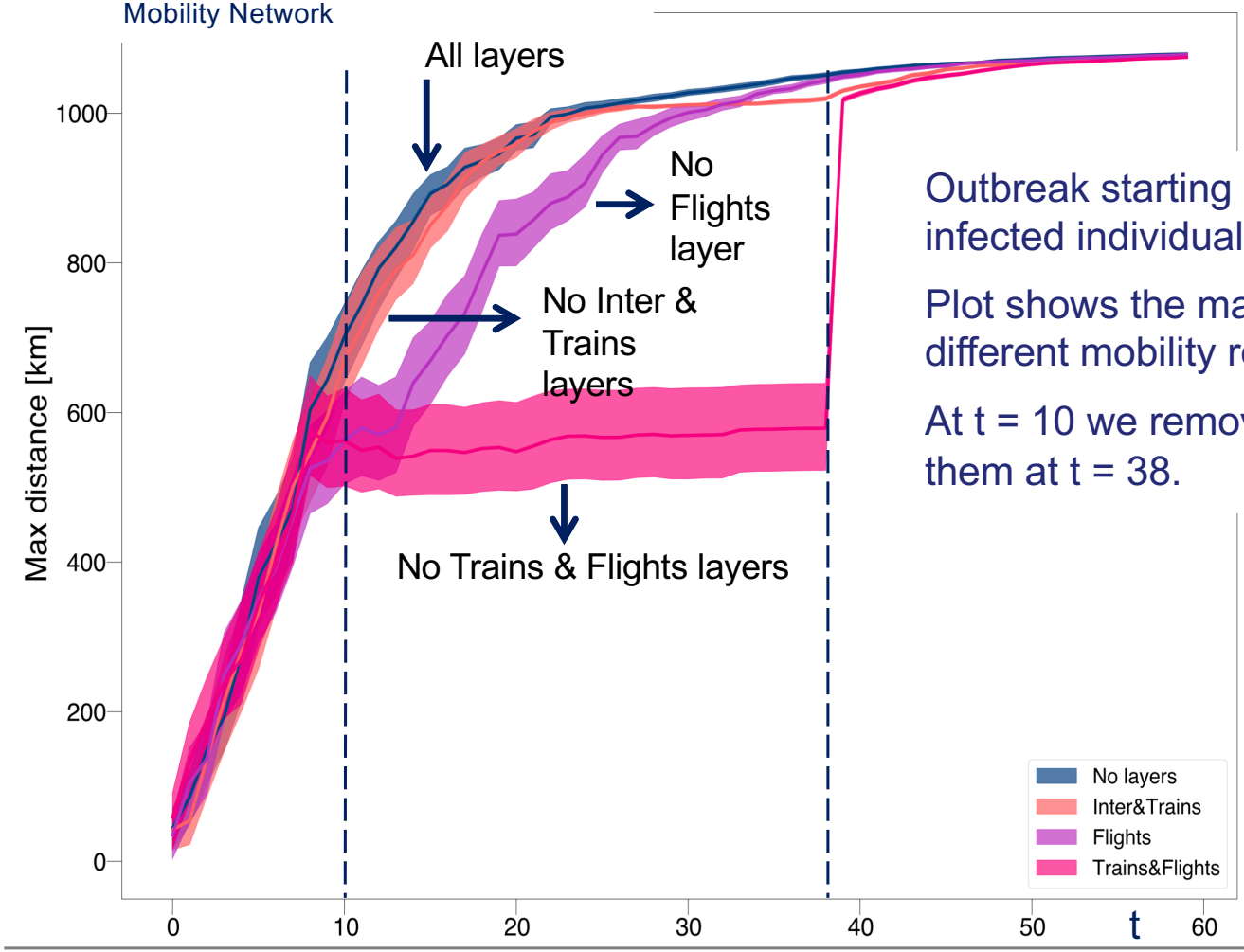
L = Flights

g_{ij} is the fraction of total travellers *i-j*. Dati ENAC



Networks Dynamics

Mobility Network



Probability that an individual moves from *i* to *j*

$$p_{ij} = p_M \sum_{\mathrm{L}} p_{\mathrm{L}} \tilde{w}_{ij}^{\mathrm{L}}$$

$\mathrm{L} \in \{\mathrm{Intra}, \mathrm{Inter}, \mathrm{Train}, \mathrm{Flight}\}$

Outbreak starting in Milan with seed size of 25 (the initially infected individuals)

Plot shows the maximum distance as a function of time for different mobility restriction scenarios.

At t = 10 we remove a single or a pair of layers and re-add them at t = 38.

Values for p_{Layer}

$p_{Intra} = 0.4$
$p_{Inter} = 0.3$
$p_{Train} = 0.2$
$p_{Flight} = 0.1$

https: //www.dati.lombardia.it/Mobilit-e-trasporti/Matrice-OD2020-Passeggeri/hyqr-mpe2

Building a Knowledge Base

Federico Zuccari, Inferno Canto III.

CovKB: A Knowledge Base for the analysis of the Covid19 epidemic in Italy

G. Salina
INFN Sezione di Roma *Tor Vergata*

G. Cimini and A. Desiderio
Physics Department, Tor Vergata University, Rome

F. Sylos Labini
Enrico Fermi Research Center, CREF

L. Vigliano
Mathematics Department, Tor Vergata University, Rome

General Framework

The outbreak of an epidemic crisis has similarities with the occurrence of an earthquake, where we do not have access to the state of the system, and thus we cannot make a precise forecast, but where we can monitor what happens and adopt adequate prevention policies. Indeed, that a global epidemic of the Covid-19 type could be triggered with a non-negligible probability was known since some decades. This is due to a series of factors, from the increase in promiscuity between humans and wild species to the development of a global interconnected world. However, the prediction of when and where an epidemic of this kind would have erupted is not possible because there are no data available to predict such an event.

Instead, it is possible to intervene promptly when the epidemic has started to spread because it is only then that it is possible to have data on its development. Computational epidemiology uses diffusion models on complex networks to make forecasts if big data concerning the number of patients, the mobility of people, etc are available. In this sense, the forecasts have a similar character to those of meteorology in which the system is observed at a certain time and the equations of fluid dynamics and thermodynamics are appropriately simplified to calculate the weather tomorrow. The combination of computing power and big data plays a fundamental role in both cases.

The problem in calculating the spread of the virus concerns the sensitive dependence on microscopic conditions and the extreme heterogeneity in its propagation. This intrinsic chaoticity in the diffusion process complicates the outlook for the epidemic a lot, especially when the diffusion is far from uniform, even when big data, computing power and mathematical models are available. The role of forecasts in the case of the spread of an epidemic is to help the political decision maker to make the most appropriate choices. The pandemic crisis cannot be avoided, but it can be avoided that it has disastrous effects with appropriate interventions, a priori and posteriori, and it is possible to change its evolution with appropriate policies. The lessons for policy makers are that of developing prevention and support systems that helps at the right time to face an epidemic. The aim of this project is that of providing a structured dataset to give a rational basis to political decision makers

Definition of a Knowledge Base (KB) for the analysis of the Covid19 epidemic in Italy (CovKB)

The defection of the CovKB will permit the integration of empirical mobility networks (short and long range) with the other subpopulation attributes (population age structure and density, health and economic infrastructures, etc.). Such a networked data structure, modulated by the timeline of

Building a Knowledge Base

The Knowledge Discovery Process (KDP)

- Selection: Capturing relevant prior knowledge, identifying the data-mining goal and developing and understanding of the application domain. Based on that, *proper data samples as well as relevant variables* can be selected.
- Pre-processing: The selected data are processed. Handling of missing values, the noise and errors identification (and correction), the elimination of duplicates, as well as the matching, fusion, and conflict resolution for data taken from different sources are done.
- Transformation: The clean dataset is transformed into a form suitable for data mining algorithms analysis. To improve the analysis performance dimensionality reduction methods can also be applied.
- Data mining:
  - Heuristic Models: the goal of the is matched to a particular method, such as classification, regression, or clustering the transformed data. A decision about which models and parameters might be appropriate must be done and matching a particular data mining method with the overall criteria of the KDP in Datasets process is necessary.
  - Theoretical Models: the transformed data are used to derive the "interaction" parameters of a theoretical dynamic model. The KDP coincides with the understanding of the system's physical dynamics.
- Evaluation and interpretation: The patterns and models derived are analysed with respect to their validity. The user assesses the usefulness of the found knowledge. A data visualization of the extracted patterns and models is involved.

The first three steps define the ETL process: these will allow one to build a solid structure for the KB, defining its fundamental variables, computational models and logical and physical architecture, i.e. the whole set of information (dataset and metadata).

This is particularly relevant, given that the analysis of the Covid19 epidemic has highlighted the lack of a solid KB useful for studying the phenomenon itself: few data available, affected by systematics that are sometimes not well defined and of poor quality.

Written two years ago. It was true then, unfortunately, it is still true today

The goal is define and implement a KB that:

- identifies the different domains and the different data sources
- is integrated and standardized between the data of the various sources, with high quality, unambiguous and homogeneous data
- is useful for studying the Covid19 in Italy and for discovering the factors that determined its evolution
- supports the simulation of different scenarios, regarding different containment strategies
- supports policymakers' decisions to be taken based on a Data-Driven approach
- is useful for defining a response strategy for any future epidemics by means of a predictive model

To reread them today, the last two points are pure Utopia

Building a Knowledge Base

The KB must contain Data and Metadata of different types:

- Concerning the parameters of the epidemic of the individuals (provinces as elementary entities)
- Concerning to the description of the spatial distribution of the elementary sets of then individuals (municipalities as elementary entities) and of all the characteristics of these elementary sets (geolocation, inhabitants, surface, urbanization level, demographic distribution, mortality data, ethnic distribution, geophysical variables, etc.)
- Concerning the structure of the National Health System (territorial location of hospitals and treatment centres and their characteristics as beds, hospitalizations, etc.). The following time series should be at a daily detail: emergency interventions and their description, clinical trials aimed at monitoring certain diseases.
- Concerning response operations aimed at containing the epidemic (government's decrees, identification of health structures aimed at curing Covid19, identification of municipalities subject to further restrictions, red areas, etc.)
- Concerning to mobility (origin destination matrices of individual movements at municipal level, road and motorway distance matrix, road and motorway flows, railway, air and naval network and respective flows)
- Concerning to the productive structure (number of industries and number of employees at the municipal level by Ateco class, matrix of goods/people flows generated by the production processes).

Epidemic Data:
- github.com/pcm-dpc/COVID-19, PL
- https://github.com/collections/open-data, ML Some
- ISS DataSet, PL Not Open

Demographic and Environmental Data:
- www.istat.it, ML
- Open Data Sites of ARPA Regional Agencies, ML
- Open Data Sites of Italian Regional Administrations, ML

National Health System Data:
- www.dati.salute.gov.it/dati/homeDataset.jsp, ML
- www.iss.it/basi-di-dati, ML
- Open Data Sites of Italian Regional Administrations, ML

Mobility Data:
- www.istat.it, ML
- dati.mit.gov.it/catalog/dataset, ML
- www.stradeanas.it/it/le-strade/osservatorio-del-traffico
- www.enac.gov.it/open-data
- dataforgood.fb.com/docs/covid19/, PL & ML
- https://news.google.com/covid19/map?hl=it&gl=IT&ceid=IT%3Ai, PL

Open Data Regional Administrations Economic Data:
- www.istat.it, ML
- www.camcom.gov.it/P43K973O0/open-data.htm, ML

Generic Data:
- www.dati.gov.it
- www.datiopen.it/it

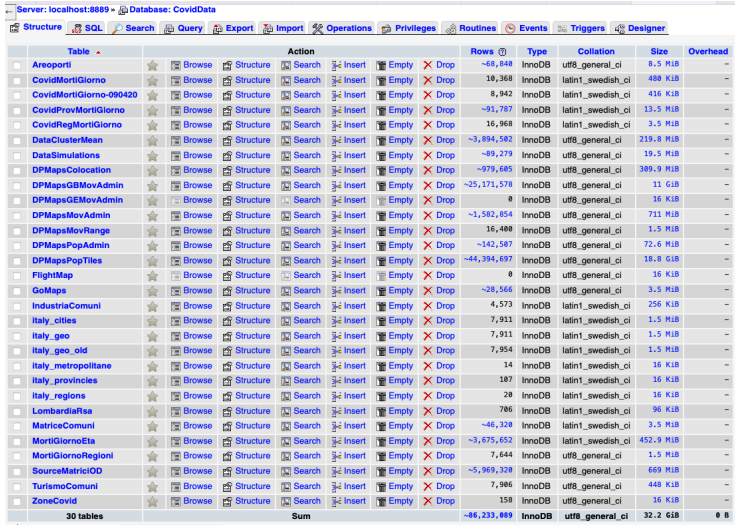
Main Open Data Resouces

Administrative Elementary Entities: ML = Municipal Level. PL = Provincial Level .

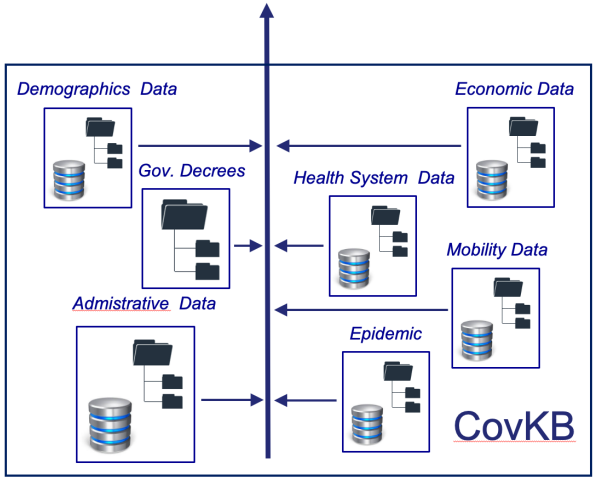
Building a Knowledge Base

The KB must contain Data and Metadata of different types:

- Concerning the parameters of the epidemic of the individuals (provinces as elementary entities)
- Concerning to the description of the spatial distribution of the elementary sets of then individuals (municipalities as elementary entities) and of all the characteristics of these elementary sets (geolocation, inhabitants, surface, urbanization level, demographic distribution, ethnic distribution, geophysical variables, etc.)
- Concerning the structure of the National Health System (territorial location of hospitals and treatment centres and their characteristics as beds, hospitalizations, etc.). The following time series should be at a daily detail: emergency interventions and their description, clinical trials aimed at monitoring certain diseases.
- Concerning response operations aimed at containing the epidemic (government's decrees, identification of health structures aimed at curing Covid19, identification of municipalities subject to further restrictions, red areas, etc.)
- Concerning to mobility (origin destination matrices of individual movements at municipal level, road and motorway distance matrix, road and motorway flows, railway, air and naval network and respective flows)
- Concerning to the productive structure (number of industries and number of employees at the municipal level by Ateco class, matrix of goods/people flows generated by the production processes).

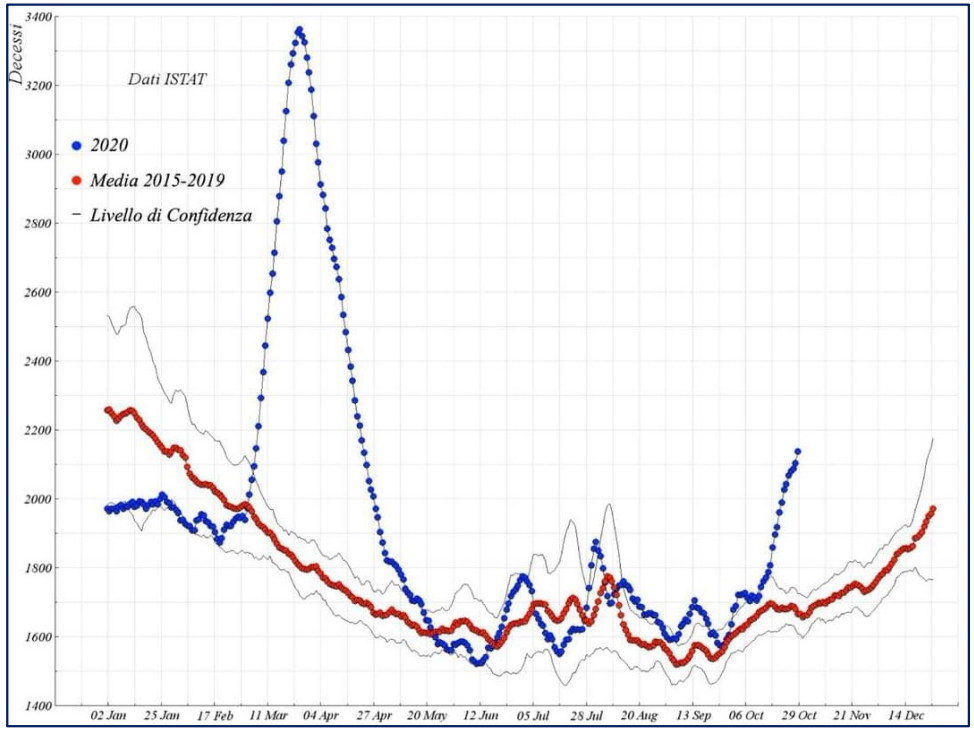


~ 90.000.000 Records

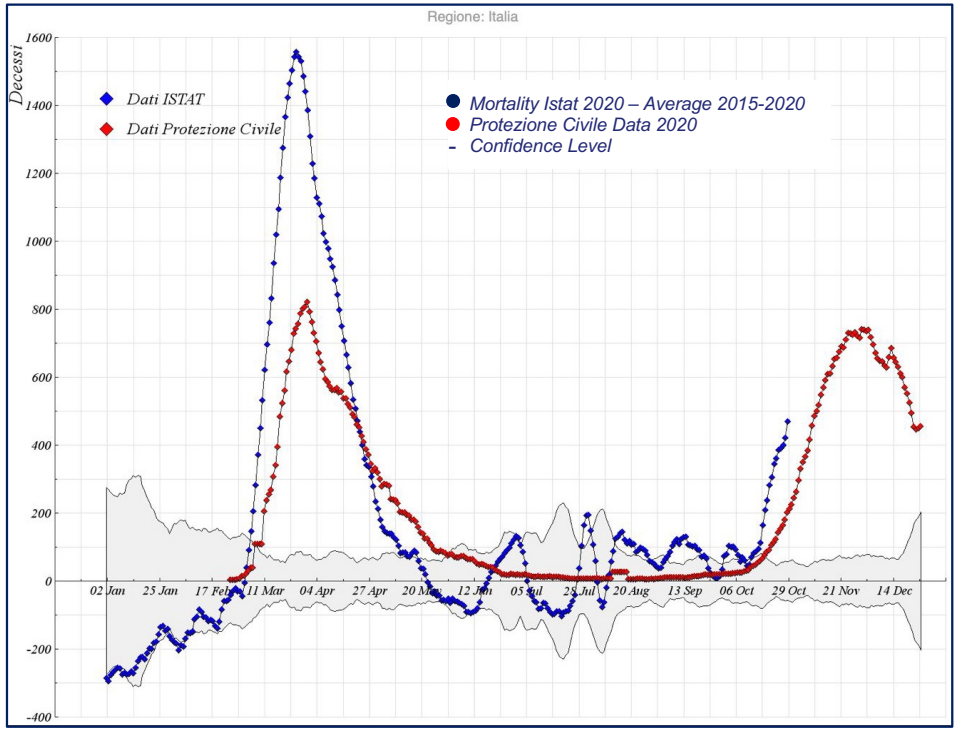


Some Analysis

Mortality March 2020 ←→ November 2020



Mortality Istat (CovKB): Source https://www.istat.it/it/archivio/240401

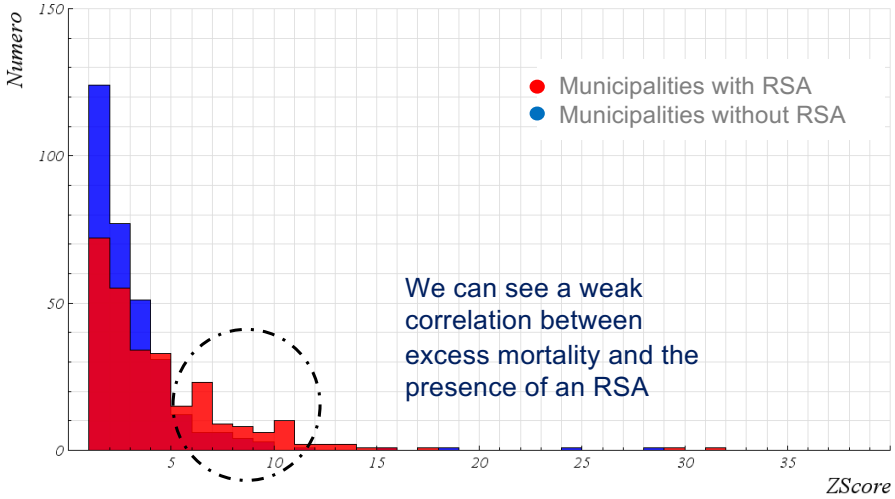


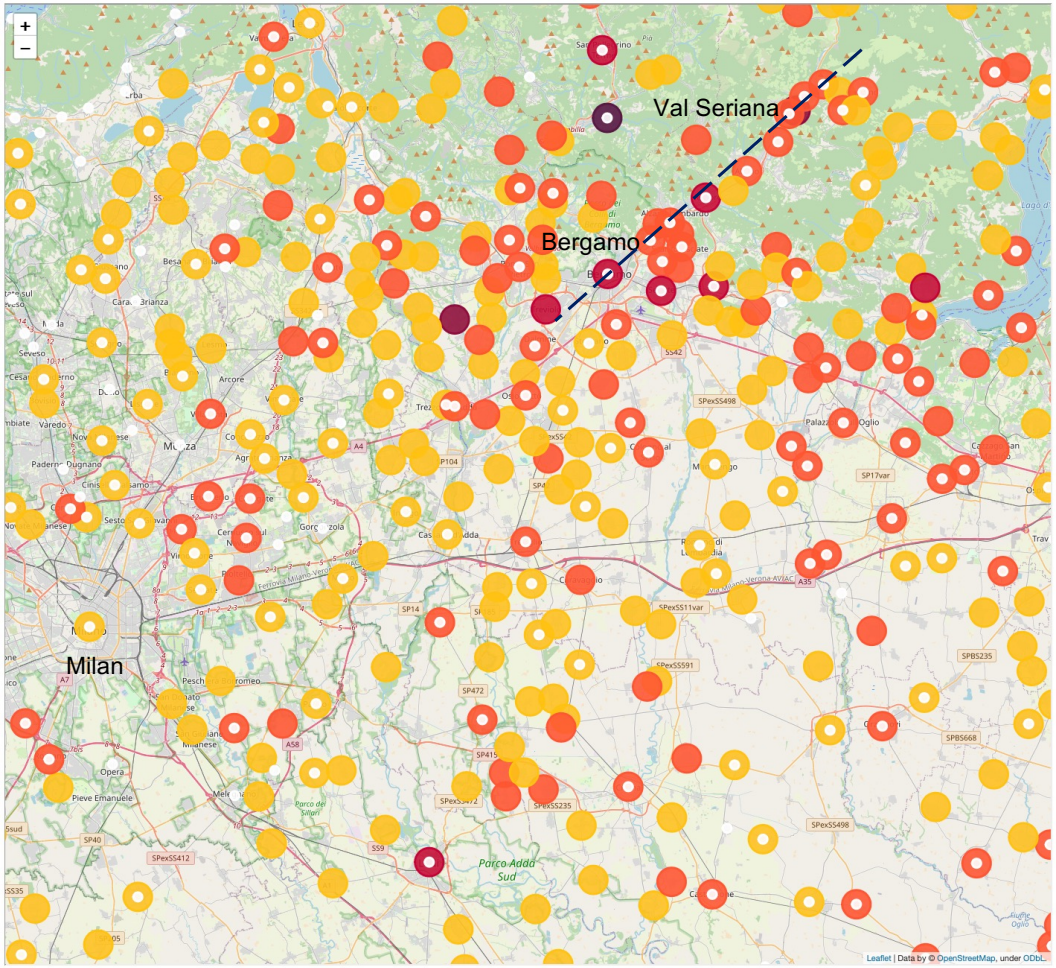
Protezione Civile Data (CovKB): Source github.com/pcm-dpc/COVID-19
Mortality Istat (CovKB): Source https://www.istat.it/it/archivio/240401

Some Analysis

Mortality & RSA: March 2020 ←→ April 2020

We analyze the excess mortality (in ZScore) in municipalities with and without RSA in Lombardy





Zscore 1

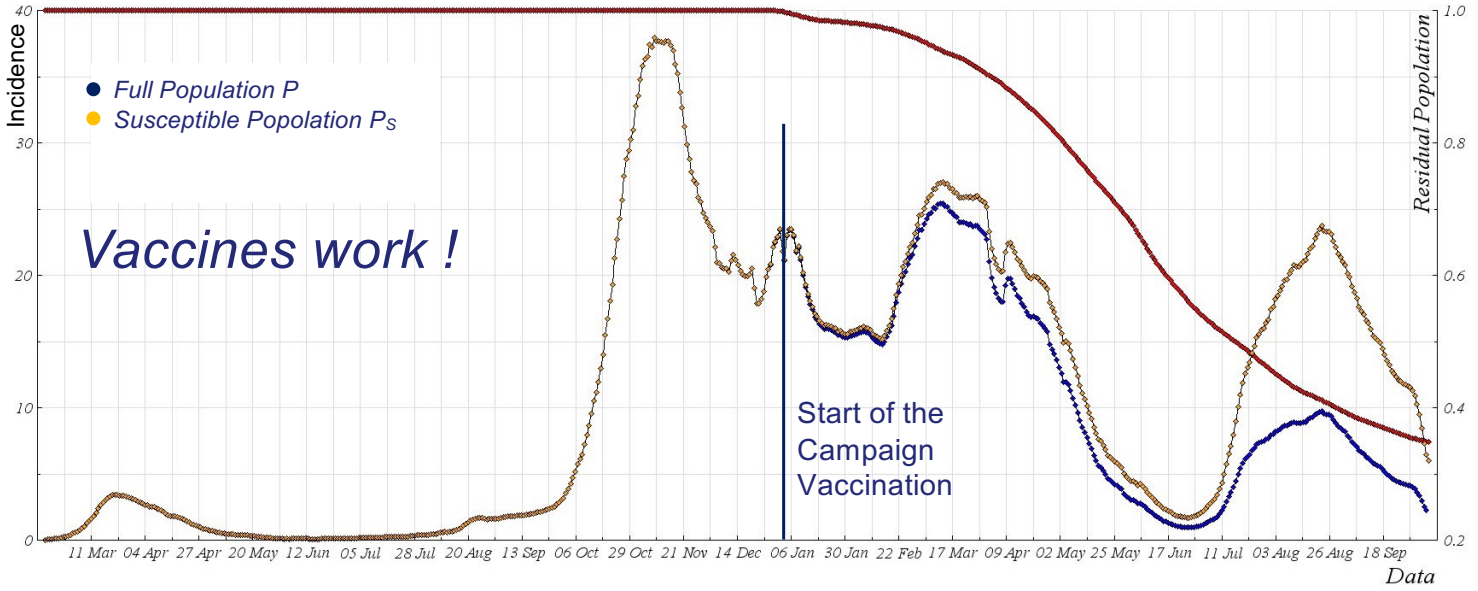
Zscore 32

RSA are Nursing home for the elderly

Some Analysis

Effect of the Vaccination Campaign: January 2020 ←→ November 2021

- In blue the incidence (number of infections per 100,000 individuals) processed by the data of the Civil Protection from March 2020 for all age groups, Full Population.
- In red (right axis) the ratio between the Susceptible Population and the Full Population. This ratio is less than one since the vaccination campaign.
- The Susceptible Population is evaluated starting from the data on the temporal evolution of vaccines. It is the sum of the number of people not vaccinated plus the fraction of the cohorts (vaccinated one dose, vaccinated two doses, vaccinated one dose + previous infection) multiplied by $1 - \alpha$, α is the average efficacy factor of the vaccines (from ISS). In orange the incidence calculated on the Susceptible Population.

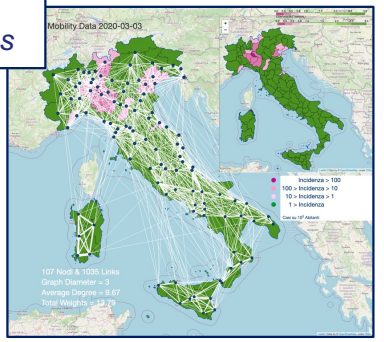


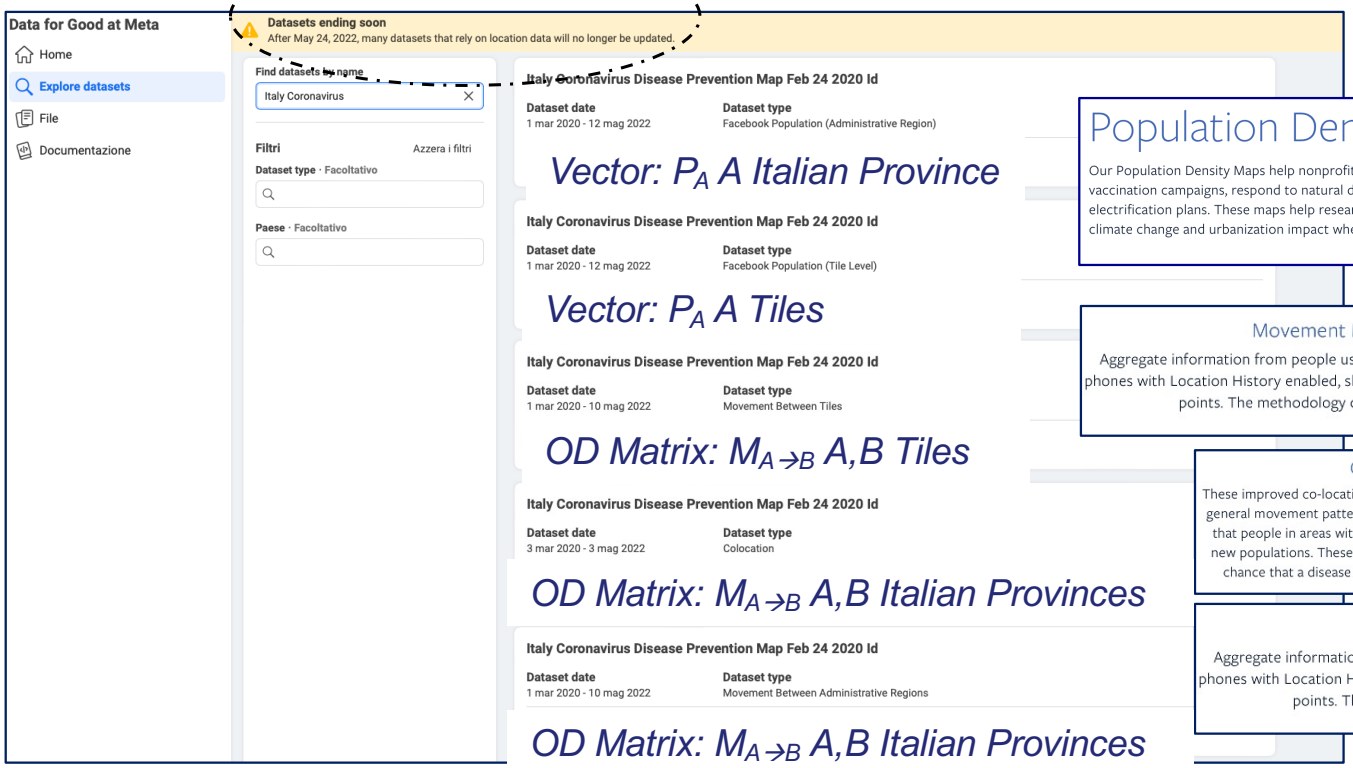
Mobility Data

FACEBOOK Data for Good

Mobility Data

Colocation Matrix: $M_{A \to B}$ A,B Italian Provinces





Population Density Maps

Our Population Density Maps help nonprofit and multilateral agencies plan vaccination campaigns, respond to natural disasters, and evaluate rural electrification plans. These maps help researchers assess the ways in which climate change and urbanization impact where people live.

Movement Maps

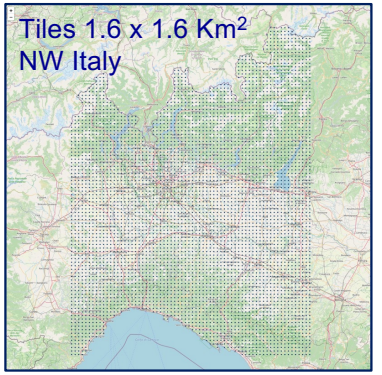
Aggregate information from people using Facebook on their mobile phones with Location History enabled, showing movement between two points. The methodology can be found here

Co-location Maps

These improved co-location insights help researchers better understand general movement patterns like commute patterns, and the probability that people in areas with disease outbreaks will come in contact with new populations. These maps are important to understand the actual chance that a disease will be spread by human-to-human contact.

Movement Maps

Aggregate information from people using Facebook on their mobile phones with Location History enabled, showing movement between two points. The methodology can be found here



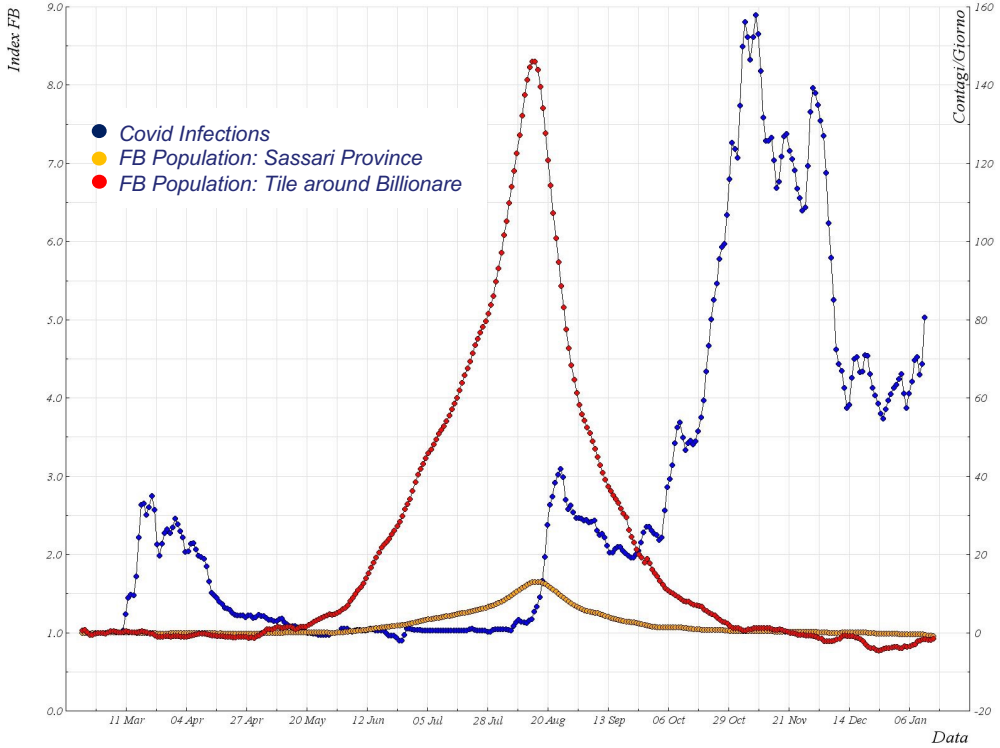
Mobility Data

Mobility Data

FACEBOOK Data for Good

Population Density Maps

Our Population Density Maps help nonprofit and multilateral agencies plan vaccination campaigns, respond to natural disasters, and evaluate rural electrification plans. These maps help researchers assess the ways in which climate change and urbanization impact where people live.



From My post on FB.

Tonight I was "lucky" to hear an interview with Flavio Briatore. He rattled off numbers with extreme ease about the economic situation. The message is always the same: "The economic crisis will kill more than the coronavirus, and that closing restaurants and nightclubs as well as being crazy is useless in order to fight the epidemic."

I would like to contrast the following graph with Briatore's numbers. In blue the infections in the province of Sassari. In orange the FB index proportional to the population present in the province of Sassari normalized to the average of the index value in the first 15 days of March. In red the same quantity in the Porto Cervo area where Il Billionare is located (it is a small square of 1.6 Km x 1.6 Km).

I don't think any comments are needed.

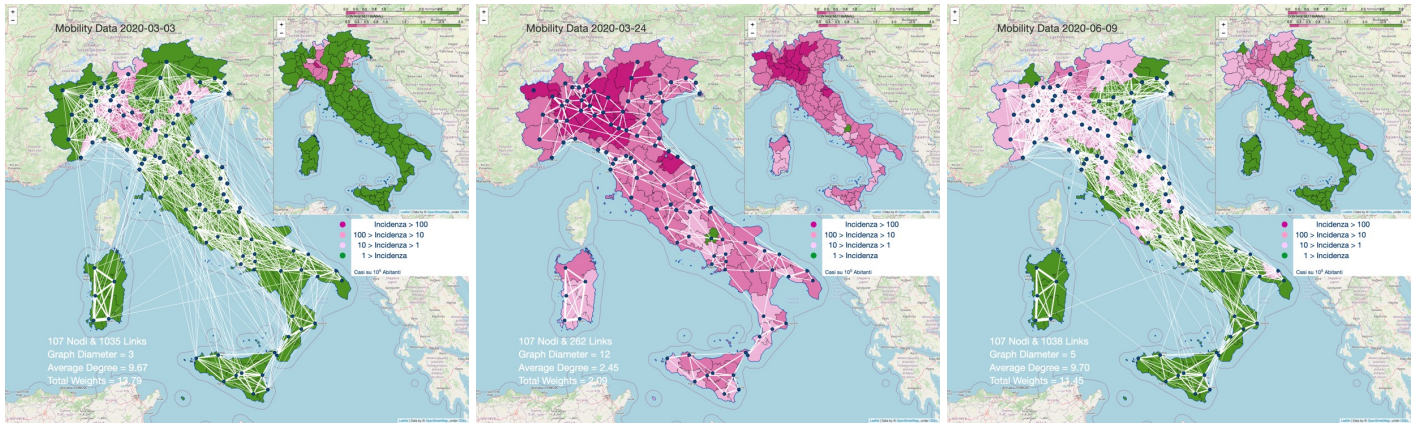
Mobility Data

FACEBOOK Data for Good

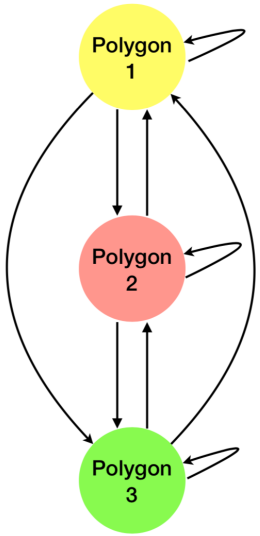
Mobility Data

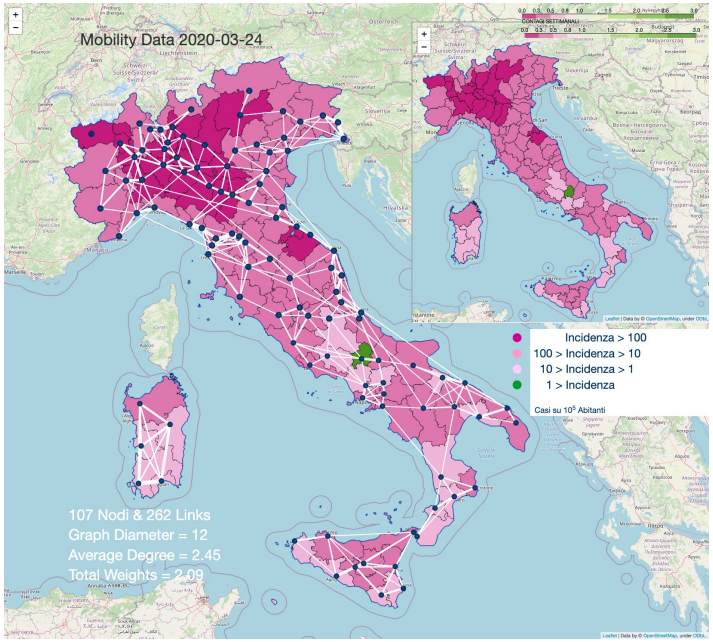
Co-location Maps

These improved co-location insights help researchers better understand general movement patterns like commute patterns, and the probability that people in areas with disease outbreaks will come in contact with new populations. These maps are important to understand the actual chance that a disease will be spread by human-to-human contact.



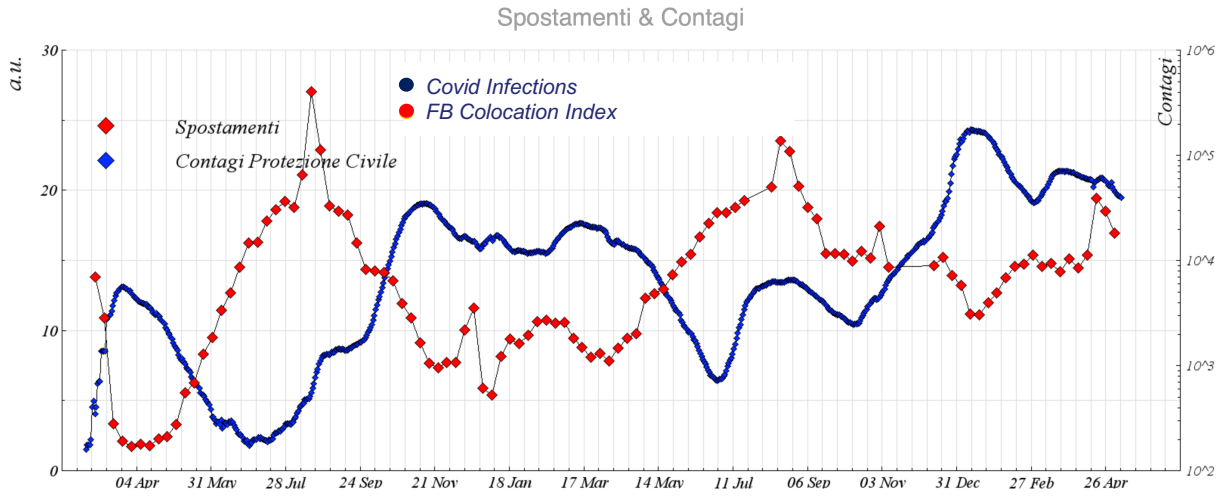
The denominator is the maximum number of colocations observed in the week. The link values are hence always symmetric as well

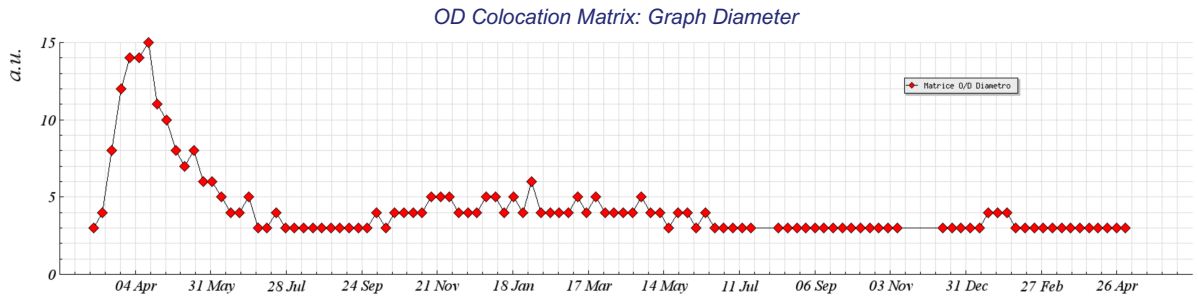


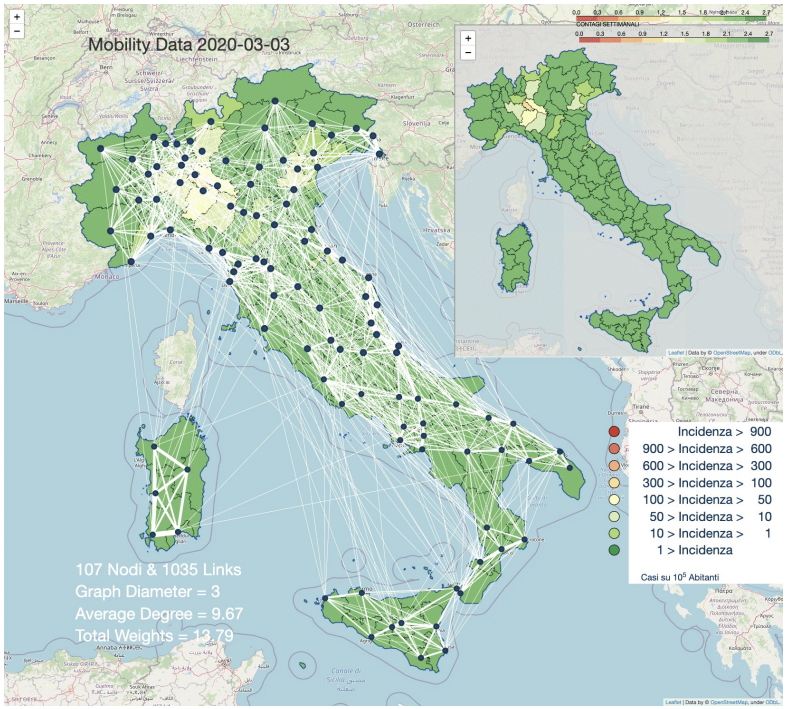


Some Analysis

Temporal evolution of the OD colocation matrix and temporal evolution of Covid infections







Human mobility and the spread of the pandemic

Some of my Considerations

- *"Few data are available, affected by systematics that are sometimes not well defined and of poor quality". It was true two years ago, remains true today.*
- *Paradoxically, in the first months of 2020 there was a greater availability of data. Movements for OpenData released epidemic data at the municipal level obtained with scraping techniques on institutional servers. When the fact was publicly known all the servers were removed from Internet.*
- *The lack of granular data for the outbreak is a purely political choice. There was no possibility of agreeing anything in the ISS-Lincei and ISS-INFN agreements. This makes the ISS dataset, in fact, no more useful than that of civil protection.*
- *Facebook offered better mobility data than those provided by the big tech companies, but their metrics do not allow easy integration with the models.*
- *You should start with the data of the telephone operators, but they are expensive and the work to arrive at the mobility data is enormous.*
- *Building a knowledge base for Covid was a beautiful illusion, but the task is bigger than the capabilities of our small group.*
- *However, the built database is very useful as a support to the study of realistic diffusion models on complex networks.*

Conclusion

SIR models on 2D Networks

+Links = 0 %

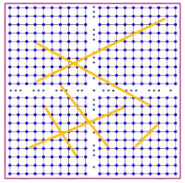
$_{ER}\lambda_c \neq {}_{NN}\lambda_c = {}_{SW}\lambda_c$

+Links = 6 %

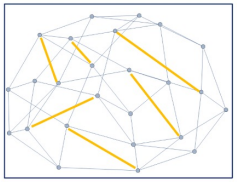
$_{ER}\lambda_c \neq {}_{NN}\lambda_c \neq {}_{SW}\lambda_c$

+Links = 50 %

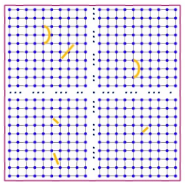
$_{ER}\lambda_c \neq {}_{NN}\lambda_c = {}_{SW}\lambda_c$



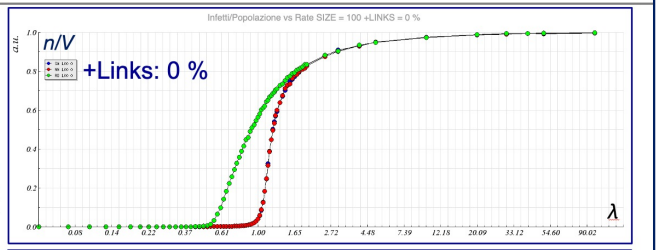
2D SW Network

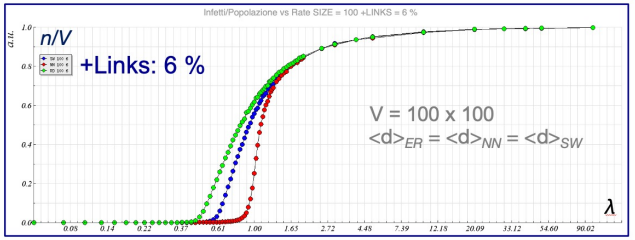


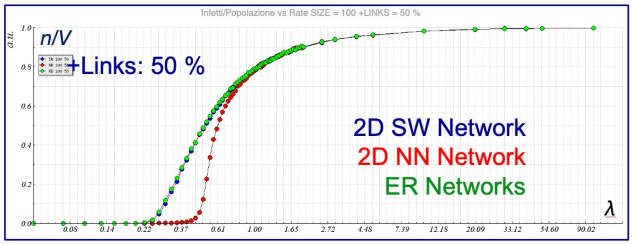
ER Network



2D NN Network







SW Networks quickly forget the 2D Square Lattice substrate.

In the limit of infinite links added, SW Networks behaves like ER Networks.

Gaetano Salina

INFN School of Statistics 2022, Paestum 17 May 2022

I go back to simulating Physical Systems on complex topology networks.

At least, in case of missing or poor quality data, I know who to contact.

Grazie !