Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Intermediate Report

Project CODE: Coupling Opinion Dynamics with Epidemics

November 30, 2024

Contents

1 Coordination, management and dissemination

1.1 Organization and financial management

The research project involves three partners:

- **CNR** with 2 full-time staff members,
- **PoliMi** with 2 full-time staff members, and
- **CREF** with 3 full-time staff members.

To strengthen the project team, CNR hired a senior researcher, **Dr. Sandro Meloni**, whose extensive expertise in the field warranted a "grant"-type research fellowship to provide a salary aligned with his level. Additionally, PoliMi recruited a junior researcher, **Dr. Fabio Mazza**, to further support the research activities.

Financial Considerations The hiring processes faced delays due to the late availability of project funds. These delays will likely necessitate reallocating resources by shifting funds from **Item A.2** to **Item A.1** of the expenditure budget. Moreover, since the project's financial rules allow reimbursement only for expenses over 500€, most of the "other costs" initially designated for travel will be reassigned to other expenditure categories. Travel costs will now be covered through alternative funding sources.

Organizational Structure The project's organizational framework includes:

- **Weekly internal meetings** within each partner institution.
- **Monthly inter-partner meetings**, conducted online, to ensure smooth collaboration.

The research activities have been coordinated by the Principal Investigator (PI) to maintain the independence of each partner while fostering the integration of their contributions to the project.

1.2 Dissemination

Website The project's website, developed and maintained by the CNR unit, is available at www.code-project.it. It provides a platform to present the project activities, introduce the partners, and publish news updates regularly.

Publications and Conferences During the first year of the project:

- A paper was accepted and will be presented at the **Complex Networks 2024** conference by **Dr. Fabio Mazza** (PoliMi).
- Several other papers were submitted to prominent journals in the field, with publication expected by spring 2025.

Dr. Sandro Meloni (CNR) was invited to give a talk at **Econets 2024**, where he highlighted the project's activities.

In January 2025, **Dr. Sandro Meloni**, **Dr. Francesca Colaiori** (CNR), and **Dr. Francesco Pierri** (PoliMi) will present project-related work at the **CS2Italy Conference**, the inaugural Italian conference on computational social science. Furthermore, abstracts describing the project's core activities have been submitted to **NetSci 2025**.

Future Plans Looking ahead, the team plans to organize a **satellite event on the behavioral aspects of epidemic containment** at **CCS 2025**, the 21st Conference of the Complex Systems Research Community.

1.3 Data and Software

Data All data collected during the project will be managed according to the project's Data Management Plan.

At the moment, we have made available, on a GitHub repository, the data necessary to create synthetic urban networks as described in the scientific report. This dataset includes, for several Italian cities:

- population geographic and age density,
- composition of households,

- distribution of voting preferences per geographic area,
- correlation between political leaning and compliance with epidemic containment strategies as estimated by previous work, and
- age-based contact matrices.

For the collection and release of data from X (formerly Twitter), we encountered challenges due to changes in X's API plans following the acquisition of the platform by Elon Musk. This will cause a delay in the release of this dataset. We are addressing this issue as follows:

To study the impact of major offline events on opinion dynamics, we focused on *Twitter/X* data related to discussions during the COVID-19 pandemic in Italy. Although *Twitter*, as it was known at the time of data collection, is not the most widely used social platform in Italy, it has played a significant role in hosting societal and political debates due to its popularity among journalists and politicians. Moreover, the details of the available data allow the study of social behavior at the user level.

After planning the analyses, **Dr. Fabio Saracco** (CREF) submitted a request via the platform's official form to access historical data in mid-March 2024. By the end of August, following multiple email exchanges, the CREF unit gained access to the X API using three separate tokens.

The main goal is to examine how discursive communities evolved during the pandemic. Due to strict download limitations, the analyses were reorganized. We used as a starting point the Twitter dataset analyzed in Caldarelli et al. 2021, selecting content creators who, between March 1 and March 8, 2020, had posted at least one tweet that was retweeted at least once. This resulted in a dataset of approximately 25k accounts.

We then downloaded all messages written by these users during a pre-pandemic period (November 1–15, 2019), yielding approximately 2.4M posts, of which nearly 500k received at least one retweet. The download of retweeters is still ongoing and has currently reached approximately 300k users.

The same download strategy was applied to messages posted by the selected users between May 1 and May 15, 2020. So far, we have downloaded 1.4M tweets, of which around 350k received at least one retweet. However, due to a bug in the code that has slowed down the process, the download of retweeters for this second period currently covers only about 200 posts.

Software The software needed to process the data and to create synthetic contact and social networks will soon be made available in the same repository that hosts the data. Additional software currently being used to analyze online social media data, simulate epidemic diffusion—including behavioral aspects in the model as described later—and analyze epidemic patterns will be released progressively as the project advances, with notifications published on the project website. Integrated software is scheduled to be released by the end of the project, in accordance with the original project plan.

2 Scientific Advancement

2.1 General Model: data-informed epidemics with media effect, opinion dynamics and behavior adoption

The main goal of this project is shedding light onto the role that the dynamics of information spreading and opinion formation play in epidemic containment. To this end, we need a general model that encompasses both the mechanisms of disease transmission – including a (possibly dynamic) network of contacts, containment measures, and individual risk perception and behaviors – and assumptions about the way each actor obtains, processes and diffuses information, to forge their own opinions and influence others.

We consider a synthetic population, whose individuals are characterized by a vector of features spanning geographic, socio-demographic and political attributes, extracted from aggregated statistics. Some of these features are used to feed a data-driven probabilistic network model, that describes physical interactions allowing diseases transmission and, possibly, word-of-mouth. Other attributes are used to parameterize individual compliance to epidemic mitigation strategies, using social media data to measure the dependence on these attributes of awareness, risk perception and policy observance. We will mostly study epidemics at the urban scale, taking advantage of fine-grained data about individual features and interaction patterns. This will also allow us to investigate whether heterogeneous patterns of attributes within a city and between cities reflect onto measurable differences in the expected patterns of epidemic diffusion.

Besides modeling and analysis, a key element of the project is the implementation of the software needed to simulate the dynamics object of the study. While promoting a data-driven approach, the software should fully support the design and analysis of hypothetical scenarios.

2.2 Population

Let lowercase indices i and j indicate individuals of the population, whereas indices r and s will be used to refer to their attributes. Each individual i is represented through a vector of attributes $\vec{x}_i = (x_{i1}, \ldots, x_{im})$. Examples of such attributes are i's age, place of residence, household, sociability, income, political affiliation, preferred sources of information (e.g., informed/misinformed), level of awareness (i.e., knowledge of the current state of the epidemics). These attribute may

- affect the way i interacts with other individuals in the physical world, i.e., the probability p_{ij} that individuals i and j interact in a way that allows the transmission of the disease or of word-of-mouth information is $p_{ij} = p(\vec{x}_i, \vec{x}_j)$ – n.b.: this type of interactions are assumed to be symmetrical, i.e., $p_{ij} = p_{ji}$;
- affect the way i interacts with other individuals in the online/digital world, i.e., the probability q_{ij} that individuals i and j interact online in a way that allows the transmission of information/opinion from i to j is $q_{ij} = q(\vec{x}_i, \vec{x}_j)$ – nb: in general, $q_{ij} \neq q_{ji}$ because i and j might not have the same influence on one another;
- affect the way i behaves during the epidemics, i.e., the attitude a_i of individual i with respect to vaccination, social-distancing, and other mitigation norms is $a_i = a(\vec{x}_i)$.

Remarks.

- Most attributes can be assumed to be *discrete/categorical*, also in consideration of the corresponding data being available in this format (e.g., density on a grid, age intervals, tax brackets, etc.). Otherwise stated, for most r, the attribute x_r is assumed to take on one of n_r possible values, i.e., $x_r \in \{\mathbf{x}_r^1, \ldots, \mathbf{x}_r^{n_r}\}$ for some $n_r > 0$.
- In most case, we do not have access to the joint distribution of the vector $\vec{x}$. In some cases we know the joint distribution of two or more attributes – e.g., how many people of a given age live in a specific area –, in some others we know the conditional distribution of one attribute with respect to another – e.g., the percentage of votes obtained by political parties in different areas –, in others we only know the marginal distribution of an attribute. In the absence of additional information, we assume that different attributes are mutually independent.

Implementation. Once the user has selected their city of interest, the current version of the software extracts the *shapefile* defining the boundary of the city territory, tesselates the territory, gets population density from WorldPop, and creates a population where each individual has a tile of residence, an age group, and a *fitness* score. The fitness is drawn from a suitable continuous distribution (e.g., Pareto or Lognormal), and can be used to control the individual propensity to establish contacts – due, for instance, to popularity, sociability or mobility. Additional scripts allow to:

- create synthetic households based on aggregated statistics about the composition of households in terms of number of individuals of different age groups, and assign the household id to each individual;
- map the tiles to voting precints, so that each individual can be given a political affiliation based on the available electoral data.

2.3 Physical network model

From here on, let us distinguish between i's fitness f_i, and all other attributes $\vec{x}_i$, which we assume being discrete, as discussed above. The finite range of values of the attributes $\vec{x}_i$ induces a partition of the population into $n = \prod_r n_r$ blocks $B_1, \ldots, B_n$. We use capital latin indices to denote the blocks, with $i \in I$ meaning that node i belongs to block B_I.

For the physical social network, we rely on the Urban Social Network (USN) model originally proposed in Guarino et al. 2021. This is a data-driven model which, in the current implementation, is built on top of the Fitness-Corrected Block Model (FCBM) defined in Bernaschi et al. 2022. Extension to other block-models is straightforward, and a possible alternative is discussed in the following. In the original formulation, the model only uses two attributes, i's coordinates on the territory and i's age, both discretized through a tessellation of the territory and the identification of a set of age intervals, to partition the population into a set of blocks defined as the combination of geographic tiles and age groups. The block-wise mixing matrix K is defined relying on previous empirical findings, that provide solid estimates for both the ratio of contacts between any two age groups, and the dependence of social links frequency upon geographic distance – typically, an inverse-power relation with exponent between 1 and 2. If synthetic households are available, the USN model maps these households to a set of *cliques*, which constitutes a separate layer in the graph. The USN can be easily extended to include multiple layers, temporal edges, or just other factors in the expression for the edge probability. For instance, once the dependence of edge frequency on some other attribute has been established – whether hypothetical or data-driven – we just need to consider said attribute in the block definition and adjust the matrix K accordingly.

FCBM. Let $K = \{K_{IJ}\}$ be a symmetric mixing matrix that quantifies how well-connected are the nodes of blocks I and J, on average. For the sake of simplicity, we assume that K is normalized so that $\sum_{IJ} K_{IJ} = 2M$, where M is the desired number of edges in the network, whereas the fitness f is normalized *per block*, such that $\sum_{i \in I} f_i = 1$. The FCBM is defined as the maximum-entropy model satisfying

$$\langle L_{IJ} \rangle_P = K_{IJ} \quad \text{for all } I, J \tag{1}$$

$$\langle \deg_i \rangle_P = f_i \sum_J K_{I_i J} \quad \text{for all } i \tag{2}$$

where L_{IJ} is the number of links between I and J, $\deg_i$ is i's degree, and $\langle \cdot \rangle$ denotes the expected value. This model constrains the expected degree of each node and the expected connectedness of each pair of blocks, while being otherwise maximally random.

RHBM. The Random Hyperbolic Block Model (RHBM) is defined as the maximum-entropy probability distribution satisfying

$$\langle E \rangle_P = E^* \tag{3}$$

$$\langle L_{IJ} \rangle_P = K_{IJ} \quad \text{for all } I, J \tag{4}$$

$$\langle \deg_i \rangle_P = f_i \sum_J K_{I_i J} \quad \text{for all } i \tag{5}$$

In addition to the expected degree sequence and mixing matrix, the formulation of the RHBM also sets the expected total energy of the system, so that the model extends both the degree-corrected stochastic block model and the $\mathbb{S}^1$ random geometric model, enjoying a combination of their properties. The model can be equivalently defined as the union of $\binom{n+1}{2}$ $\mathbb{S}^1$ graphs, corresponding to the subgraphs induced by intra- and inter-block connections of the n communities **guarino2023hyperbolic**. In this case, the latent features of each node must be extracted once and for all, so that the centrality and homophily patterns are preserved across different subgraphs. With respect to the FCBM, the model introduces an element of node similarity by means of a latent geometry, in such a way to obtain a non-vanishing clustering (i.e., transitivity) that can be adjusted to mimic that of real world networks. If the angular coordinates θ_i's associated to each node are taken uniformly at random in $[0, 2\pi)$, the edge probability takes the form

$$p_{ij} = \frac{1}{1 + \left(\frac{\Delta(\theta_i, \theta_j) \mathcal{I}_\beta}{\pi K_{IJ} f_i f_j} \right)^\beta} \tag{6}$$

where the inverse temperature β controls the clustering of the network, and $\mathcal{I}_\beta$ is a constant that only depends on β.

Implementation. The current implementation of the USN is based on the FCBM, and it supports both the exact formulation and an approximate formulation that is significantly more efficient. An implementation of the RHBM is ready to use, but it is yet to be incorporated in the USN software.

2.4 Epidemic model

Several variations of the standard compartmental epidemic models have been proposed in the literature to account for complex disease dynamics and individual behavior. Since we need to model compliance to mitigation policies, we will focus on models that include at least one of the following: (i) heterogeneous susceptibility/infectiousness; (ii) vaccines, with individuals having variable attitude towards the vaccination campaign. The behavior/riskiness of an individual will be dependent on some of their attributes, either directly or indirectly.

2.4.1 Heterogeneous behavior/transmission

At first, we assume that the two dynamics of information and epidemic spreading do not occur concurrently, but that the former precedes the latter. Under this working hypothesis, the behavior of the population is *time invariable* and we can focus on behavior heterogeneity at the individual level. The analysis consists in comparing a benchmark scenario, where the entire population behaves in the same way, with test cases where some subsets of *misinformed/non-compliant* and/or *well-informed/law-abiding* individuals take on behaviors that are, respectively, more dangerous and safer than the average.

To this end, we introduced a model within the Susceptible-Infected-Removed (SIR) class that accounts for these heterogeneities by means of two parameters a and b that control the increasing infectivity and susceptibility of Misbehaving individuals (M) with respect to Ordinary ones (O). We found that there is a region in the space of parameters just above the epidemic threshold, where trajectories showing an initial decline in the number of Infected can suddenly reverse and give rise to widespread transmission. Such heterogeneity can lead to an underestimation of transmission potential and delayed recognition of epidemic resurgence, thereby severely compromising efforts for a timely response. We examined this phenomenon in the mean-field scenario and then simulate the dynamics on homogeneous and heterogeneous contact networks, confirming that this phenomenology persists beyond mean field. A research paper describing the model and the first set of analytical and experimental results will be presented at Complex Networks 2024.

We already started working on an extension to this model where the probability of non-compliance is correlated with individual attributes such as political affiliation, income, mobility or age. As a first test case, we have taken electoral data divided by voting precint, so that different areas of the territory are mapped to a different political landscape, and considered two approaches:

- Defining hypothetical scenarios, with a part of the political spectrum being favorable to the mitigation policies and the rest being against – e.g., supporters of the government *vs.* supporters of the opposition.
- Using social media data to estimate the relation between the political affiliation and: (i) the attitude towards a specific mitigation policy; or (ii) the exposure to good/bad-quality information, which, in turn, we assume to be a main driver of policy compliance.

The final goal is understanding whether socio-demographic and geographic differences between and within cities result in different epidemic trends when exposed to propaganda.

Implementation. Both models have already been implemented in Julia. The Julia code has been written to account for a custom-defined epidemic model, with transition probabilities that can be assigned on state changes and individually. The only fixed assumption of the Julia code is that the time is discretized.

Time variant extention To capture risk perception and possible behavioral changes among at-risk individuals, we couple disease spreading with social dynamics. Specifically, misbehaving individuals may begin to adhere to safety measures, thereby becoming ordinary, if their perceived risk of infection exceeds a certain threshold, θ. Individual perceived risk is modeled as the fraction of an individual's contacts within the network that are currently infected. This leads to a complex contagion dynamic, where social behavior –i.e., adherence to guidelines– is influenced by the local prevalence of the disease. Although our model is minimal by design, it captures the delicate interplay between disease spread and human behavioral changes. For instance, a large local prevalence will lead to a decrease in the size of the misbehaving population, thus hindering the spread. However, disease circulation is needed to increase perceived risk and convince misbehaving individuals to transition. The results of numerical simulations and theoretical analysis confirm this behavior, highlighting various effects that could have important repercussions for social system modeling and policymaking. This work will be presented at the CS2Italy Conference in early 2025.

2.4.2 Vaccination and fading immunity

Vaccination is, historically, a vastly debated topic with strongly polarized positions. While we are not necessarily interested in studying a model of the recent COVID pandemic, the paradigm change in the processes of information spreading and opinion formation triggered by the diffusion of online social media

is especially relevant when the sudden rise of cases leads to the fast development of vaccines and when the immunity – whether from natural infection or vaccine-induced – is only partial and fading in time. We therefore considered a modified-SIS model where users can be vaccinated, possibly with multiple doses, but sooner or later they return susceptible. The first version of the model ignores vaccine hesitancy, with the goal of understanding whether an optimal inter-dose period –or, equivalently, an optimal two-doses roll-out– exists. This model also allows estimating the penalty caused to the population, with respect to the optimal scenario, when the roll-out cannot be optimized due to a limited collaboration from the population. Based on a simple model with just 3 levels of susceptivity and 2 levels of infectivity, it can be shown that a non-trivial optimal vaccination rollout exists, provided that the vaccination capacity exceeds a precise threshold —under which only "first doses" should be distributed. We plan to submit this work to APS Physical Review Letters at the beginning of 2025.

Implementation. An implementation based on the Python library Epidemics On Networks (EON) has been used for simulations that include a Quasy-Stationary (QS) method to estimate the epidemic prevalence and threshold.

2.5 Online social networks

The network of online interactions occurring on X/Twitter is strongly clustered and can be effectively partitioned by means of standard descriptive community detection algorithms (e.g., modularity-based). However, off-the-shelf algorithms are generally produce a multitude of small communities that are hard to classify.

To address this issue, we thoroughly analyzed the approach originally proposed in Becatti et al. 2019 to identify so-called "discursive" communities. We came up with two different versions of the algorithm that differ in how the ideological core of such communities is identified: in the MonoDC (*Monopartite-network-induced Discursive Communities*) we consider the direct interactions –i.e. retweets– between a set of content creators; in the BiDC (*Bipartite-network-induced Discursive Communities*) we instead build a similarity network based on the audience shared by any two such users.

These algorithms provides a neater partition of the network in just a few, easy to interpret, communities, highlighting how these communities emerge mainly as a consequence of the activity of content creators, and of statistically significant patterns of preference in the retweeting habits of their most responsive audience. The core of these communities is composed of verified users – whose political affiliation is well-known – and the audience members recognize content creators whose stance on the debated topic they consider affine to their own opinions and belief systems. This leads to an interesting internal structure of these communities, which can be highlighted by a standard bow-tie decomposition: when compared with a randomized null model, we observe surprisingly small IN and SCC components, compensated by larger than expected IN-TENDRILS and OUT components. In other words, opinions are mostly produced by just a few users (IN+SCC) and reach the great majority either directly (IN→IN-TENDRILS and SCC→OUT) or through a reorganization of thoughts operated by a small set of opinion leaders (IN→SCC→OUT).

This suggests to proceed as follows, based on the purpose of the analysis:

- Collect Twitter data from the COVID-19 pandemic and use BiDC to find and characterize communities, to detect the dominant opinion about containment measures in different communities.
- Use a simple block model – such as the FCBM or the RHBM model described above – to generate synthetic online social networks, tuning the size and connectedness of the different clusters based on real data.

- To simulate information spreading, define a directed variant of the model, so that the bow-tie decomposition of the obtained synthetic network statistically resembles the measured one.

Possible ways to combine this model for online interactions with the previously described contact network models and epidemic models include:

- Studying how the dynamics of information spreading change if individuals are more likely to spread anti-establishment propaganda when the situation described by the media differs from the one they experience in their surrounding.
- Assume that both sources of true and false information exist, and behaviors are updated based on the quality of the received information; compare different scenarios in terms of the structure of the online network (e.g., good/bad sources are clustered or well-mixed) and its dependence on individual attributes (e.g., membership to good/bad cluster depends on political affiliation, income, age, etc.).
- Repeat the aforementiond analyses, comparing cases where the online and the interaction networks have different levels of correlation.

Implementation The MonoDC and BiDC algorithms are both implemented and ready to use, as well as the block models presented in subsection 2.3. The extension to directed networks that preserve the observed structure is instead on-going work.

References

Becatti, Carolina et al. (Dec. 2019). "Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections". In: *Palgrave Communications* 5 (1), pp. 1–16. ISSN: 20551045. DOI: 10.1057/s41599-019-0300-3.

Bernaschi, Massimo et al. (2022). "The Fitness-Corrected Block Model, or how to create maximum-entropy data-driven spatial social networks". In: *Scientific Reports* 12.1, p. 18206.

Caldarelli, Guido et al. (July 2021). "Flow of online misinformation during the peak of the COVID-19 pandemic in Italy". In: *EPJ Data Science 2021 10:1* 10 (1), pp. 1–23. ISSN: 2193-1127. DOI: 10.1140/EPJDS/S13688-021-00289-4. URL: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00289-4.

Guarino, Stefano et al. (2021). "Inferring Urban Social Networks from Publicly Available Data". In: *Future Internet* 13.5. ISSN: 1999-5903. DOI: 10.3390/fi13050108. URL: https://www.mdpi.com/1999-5903/13/5/108.