NetSci 2023

International School and Conference on Network Science

7-10 February 2023
Buenos Aires, Argentina

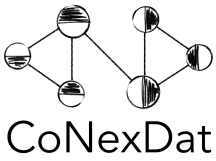

NetSci


.UBA fiuba
FACULTAD DE INGENIERÍA


FRANCE 2030
Labex MME-DII
Modèles Mathématiques et Économiques de la Dynamique, de l'Incertitude et des Interactions


CY
ADVANCED STUDIES


.UBA EXACTAS
SED
Discrete
Event
Simulation Lab


CoNexDat


.UBA INVESTIGACIÓN
Secretaría de Ciencia y Técnica

A graph complexity measure based on the spectral analysis of the Laplace operator

Diego M. Mateos [1,2,3*], Federico Morana[3], and Hugo Aimar[1,3]

[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.
[2]Facultad de Ciencia y Tecnología. Universidad Autónoma de Entre Ríos (UADER). Oro Verde, Entre Ríos, Argentina.
[3]Instituto de Matemática Aplicada del Litoral-UNL-CONICET (IMAL), CCT CONICET, Santa Fé, Argentina.
[*]Corresponding author: Diego M. Mateos, mateosdiego@gmail.com.

October 25, 2022

Graph, Complexity, Laplacian, Spectral analysis

1 Introduction

In this work we introduce a concept of complexity for undirected graphs in terms of the spectral analysis of the Laplacian operator defined by the incidence matrix of the graph. Precisely, we compute the norm of the vector of eigenvalues of both the graph and its complement and take their product. Doing so, we obtain a quantity that satisfies two basic properties that are the expected for a measure of complexity. First, complexity of fully connected and fully disconnected graphs vanish. Second, complexity of complementary graphs coincide. This notion of complexity allows us to distinguish different kinds of graphs by placing them in a "croissant-shaped" region of the plane link density - complexity, highlighting some features like connectivity, concentration, uniformity or regularity and existence of clique-like clusters. Indeed, considering graphs with a fixed number of nodes, by plotting the link density versus the complexity we find that graphs generated by different methods take place at different regions of the plane.

2 The mathematical model

Let $G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ be a simple undirected graph, where $\mathcal{V} = \{1, \dots, n\}$ is the set of vertices or nodes, $\mathcal{E} = \{e_1, \dots, e_m\} \subset \{\{i, j\} \colon i, j \in \mathcal{V}\}$ is the set of edges and $\mathcal{W} \colon \mathcal{V} \times \mathcal{V} \to \{0, 1\}$ is the adjacency matrix of G with $w_{ij} = 1$ whenever $\{i, j\} \in \mathcal{E}$ and zero otherwise. Since the graph is undirected and simple the matrix $\mathcal{W}$ is symmetric with null diagonal. We will denote $i \sim j$ when $\{i, j\} \in \mathcal{E}$.

The degree of a vertex j is defined by $\delta(j) = \sum_{i \in \mathcal{V}} w_{ij}$. The degree matrix is defined as the diagonal $n \times n$ matrix containing the degrees of the nodes and denoted by $D = diag(\delta(1), \dots, \delta(n))$. The Laplacian of the graph is the lineal operator acting on real or complex functions defined on the nodes, with matrix given by

$$\Delta = \mathcal{W} - D. \tag{1}$$

This operator is symmetric and negative semi-definite. Therefore we can apply the spectral theorem to obtain an orthonormal basis of $\ell^2(\mathcal{V}) \sim \mathbb{R}^n$ of eigenvectors $\{\psi_1, \dots, \psi_n\}$ of Δ. It is usually called the Fourier basis of G. The associated eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ satisfy $0 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. In the following we will refer to the vector $\overline{\lambda} = (\lambda_1, \dots, \lambda_n)$ as the spectrum of the graph G. The trace of the Laplacian is a feature of interest for our further analysis and is given by $\sum_{i=1}^{n} \lambda_i = -2m$, where m is the number of edges of G. For a general reference regarding the spectral theory of the Laplacian on graphs see [1] and references therein.

We may consider equivalent two graphs G and G' that share the spectrum $\bar{\lambda}$. Hence, for G and H two graphs with the same number n of vertices, the function $d_s(G, H) = \left\| \bar{\lambda}_G - \bar{\lambda}_H \right\|$, with $\bar{\lambda}_G$ and $\bar{\lambda}_H$ the spectrum of G and H respectively and $\|\cdot\|$ any norm in $\mathbb{R}^n$, is a distance (metric) between the classes of co-spectrality of G and H. We shall take the usual (euclidean) norm $\left\| \bar{\lambda} \right\| = \left(\sum_{i=1}^{n} |\lambda_i|^2 \right)^{1/2}$. We shall refer to d_s as the spectral distance. Notice that since the first eigenvalue λ_1 of each graph vanishes, we actually have that $d_s(G, H) = |\bar{\Lambda}_G - \bar{\Lambda}_H|$,

where $\bar{\Lambda} = (\lambda_2, ..., \lambda_n)$ and $|\cdot|$ is the euclidean norm in $\mathbb{R}^{n-1}$. The spectral distance on graphs was considered before in [2], see also [3].

In order to introduce our definition of spectral complexity of a graph, let us set Z to denote the null graph, i.e. $w_{ij} = 0$ for every $i, j \in \mathcal{V}$, and F to denote the complete graph, i.e. $w_{ij} = 1$ for every $i \neq j$. Now we can define the **spectral complexity** of a graph G with n vertices as

$$\begin{aligned} \mathcal{C}_s(G) &= d_s(G, Z) \cdot d_s(G, F) \\ &= \left\|\bar{\lambda}_G - \bar{\lambda}_Z\right\| \ \left\|\bar{\lambda}_G - \bar{\lambda}_F\right\| \\ &= |\bar{\Lambda}_G - \bar{\Lambda}_Z| \ |\bar{\Lambda}_G - \bar{\Lambda}_F|. \end{aligned} \tag{2}$$

Two basic premises are behind this definition. The first one is that both, the null graph and the full graph, are the less complex graphs that can be defined on the vertices set $\mathcal{V} = \{1, ..., n\}$. The second is that complementary graphs should have the same complexity.

A second quantity associated to a graph that we shall take into account in our analysis is its link density. The link density ρ of a simple unidirected graph is the number of actual edges divided by the number of all possible edges. With our notation

$$\rho(G) = \frac{2m}{n(n-1)}. \tag{3}$$

Given a positive integer n we shall display all the possible graphs G built on $\mathcal{V} = \{1, 2, ..., n\}$ in the plane of the variables $\rho(G)$ and $\mathcal{C}_s(G)$. Since the density of a graph G and the density of its complement can be quite different, actually $\rho(G) + \rho(G^c) = 1$, it is clear that the link density is not a function of the spectral complexity. It is also simple to show that graphs with the same density may have different spectral complexity. So neither ρ is a function of $\mathcal{C}_s$ nor $\mathcal{C}_s$ is a function of ρ. As could be expected. Nevertheless ρ and $\mathcal{C}_s$ are not completely independent. In fact we empirically determine the region in the region in the plane $(\rho, \mathcal{C}_s)$ spanned by all possible graphs.

The delimitation of the region in the representation plane link *density - complexity* where all the variety of graphs take place is not a trivial task to perform theoretically. Here we obtain an empirical approximation of the upper and lower boundaries, derived after placing a wide variety of graphs generated by random and deterministic methods. The plane obtained is a croissant-shaped. Figure 1 depicts the croissant-shaped and the placement of some paradigmatic graphs of 15 vertices. This notion of plane highlight some features like connectivity, concentration, uniformity or regularity and existence of clique-like clusters

3 Result

Using the plane ρ *vs* $\mathcal{C}_s$, we analysed three well known stochastic models, the Erdös-Rényi model [4], the Watts-Strogatz model [5], and the Barabási-Albert model [6] for the different parameter each one. As we shall see, each of them draws some characteristic pattern contained in the basic croissant shape (see Figure 1b). Finally, as an application to graphs generated by real measurements, we consider the brain connectivity graphs from two epileptic patients obtained from magnetoencephalography (MEG) recording, both in a baseline period and in ictal periods (epileptic seizures). In this case, our definition of complexity could be used as a tool for discerning between states, by the analysis of differences at distinct frequencies of the MEG recording.

References

[1] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

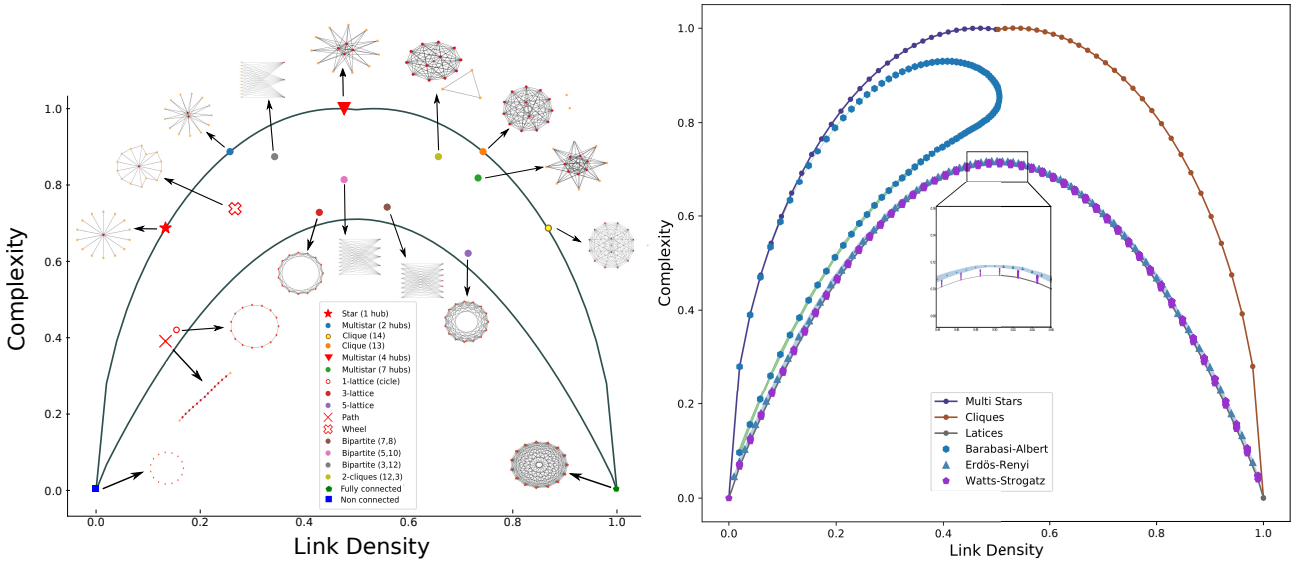
[2] Jiao Gu, Bobo Hua, and Shiping Liu. Spectral distances on graphs. *Discrete Applied Mathematics*, 190:56–74, 2015.

[3] M.M. Deza and E. Deza. *Encyclopedia of Distances*. Springer-Verlag Berlin Heidelberg, 4th edition, 2016.

[4] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[5] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[6] A-L Barabási and R Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.



(a) (b)

Figure 1: Spectral complexity vs link density plane. The full line represent the lower and upper limit. (a) Schematic distribution of the different types of networks for n = 15 on the "croissant-shaped" region. (b) Overview and comparison of the results obtained for all the network models analysed in this work. In this case we use for all models n = 100

A Manifold Minimization Principle for Physical Networks

Xiangyi Meng,[1] Csaba Both,[1] Baruch Barzel,[1,2] and Albert-László Barabási[1,3,4]

[1] *Network Science Institute and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA*
[2] *Department of Mathematics, Bar-Ilan University, Ramat Gan, 5290002, Israel*
[3] *Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA*
[4] *Department of Network and Data Science, Central European University, Budapest 1051, Hungary*

Keywords: Physical network, Manifold, Biological morphology

Physical networks [1], such as blood vessels, corals, plants, and neurons, are network-based physical objects that notably differ from abstract networks. Not only that physical networks typically have different traditional network metrics (such as distributions of degrees and cycles) than abstract networks [2], they are also endowed with rich geometry, such as distance, crossing angle, curvature, etc., that cannot be described by adjacency matrix only [1]. Despite an enormous body of studies on the *principle of formation* of abstract networks, e.g., the preferential attachment [3] or the fitness [4] mechanisms, there is not much literature on the principle of formation of physical networks, by which the nodes and links, rather than *a priori* defined, should naturally emerge from some geometric principle.

One such principle is wiring minimization, a.k.a. the Steiner problem [5], aiming at linking all destinations (terminals) with minimum total link length, which has long been considered as the physical interpretation of emerging network structures in complex systems such as neurons [6] and ant tunnels [7]. Promisingly, the solution of the Steiner problem naturally induces a tree-network structure, which nonetheless predicts the following geometric characteristics: (1) the tree only has degree ≤ 3 (i.e., bifurcation only); (2) all bifurcation angles are 120 degree (except at terminals); (3) all bifurcations are planar (even in higher-dimensional ambient space). By contrast, the Steiner predictions (1) and (2) are frequently violated in real physical networks. For example, nodes of degree 4 (trifurcation) cannot appear in the Steiner solution, since any degree-4 node can be locally replaced by two degree-3 nodes that produce a shorter overall link length, yet trifurcations are frequently observed in different biological systems [8]. Also, while most bifurcations are planar [9], the bifurcating branches are found usually not crossing at 120 degree, but many times even at 180 degree, a "sprouting" behavior that is frequently observed in, for example, blood vessels (sprouting angiogenesis) [10].

This prompts us to look into other candidates as the principle of formation of physical networks. Recent advance in physical networks indicates that treating links as having shapes leads to completely different geometric characteristics than as shapeless wires [1]. We are thus motivated to promote a graph to a higher-dimensional geometric object, namely, a smooth manifold [11]—a topological space that is everywhere locally similar enough to some Euclidean space, where we can define calculus and calculate geometric quantities. This leads to a principle of formation on $d \geq 2$ dimensions—a *manifold minimization principle* that, as we will see, gives rise to a tree-network solution that is analogous to the Steiner solution, yet need not follow the Steiner predictions (1) and (2), in full accordance with observations in real-world physical networks.

Manifold minimization principle.—Here, we focus on $d = 2$ manifolds, i.e., surfaces. *Surface area minimization* (Plateau's problem) has been extensively studied, which seems a natural generalization of the 1D Steiner problem if we can fix terminals as boundaries and study the minimal surface connecting them [Fig. 1(a)]. Unfortunately, the existence of long physical links is forbidden in Plateau's problem. Indeed, if we fix two parallel and identical circles as boundaries, then the minimal surface that connects the boundaries is a physically disconnected Goldschmidt solution [14] when $d/w > 0.168$ [Fig. 1(a)], where d is the distance between the two circles and w is the circumference of each circle. As a comparison, physical links in a real biological network typically have a much larger ratio $d/w \approx 10^0 \sim 10^1$. This indicates that real physical networks do not follow a surface minimization principle based merely on minimizing the area with no other constraints.

A key feature of physical networks is that the links must maintain *transportational functionality*: it is a necessary condition for physical links to have a physically connected skin in order to transport nutrients (e.g., tree bark) or signals (e.g., neuronal membrane). This prompts us to use the length scale w as a constraint and consider a systolic surface minimization problem: we require that every *systole* of the surface, defined as the shortest closed curve(s) on the surface that cannot be continuously contracted to a point because of topological holes it essentially surrounds, must have length w [Fig. 1(a)].

In the case of the circles as boundaries [Fig. 1(a)], the cylinder is a trivial solution to the systolic surface minimization problem. The surface maintains transporta-

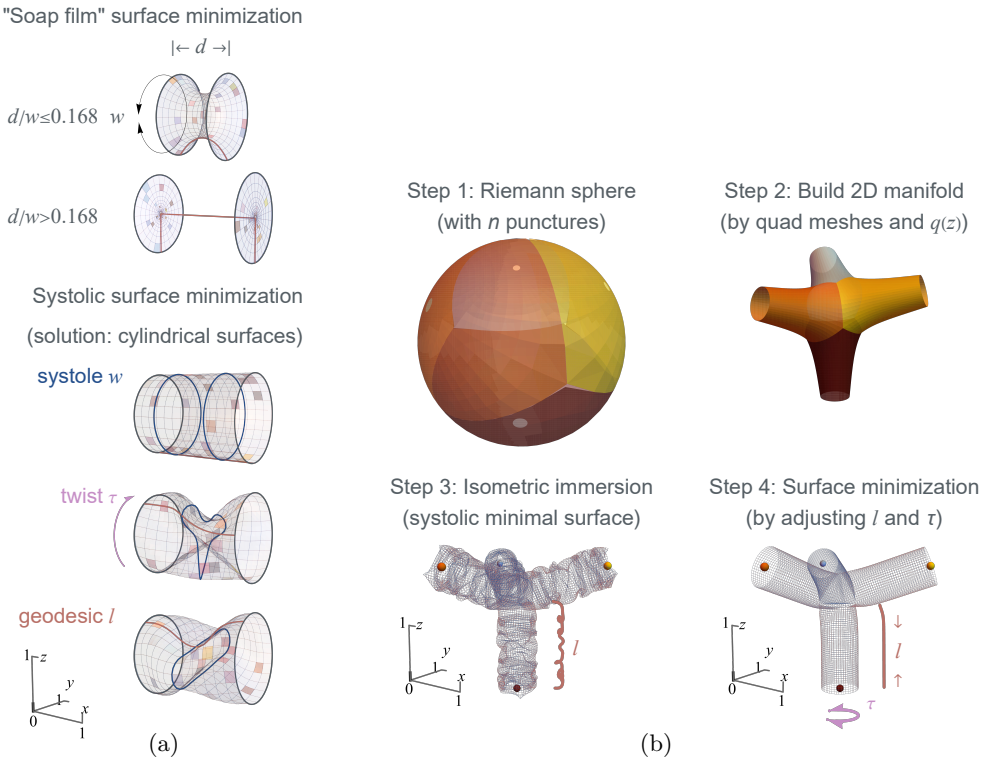


FIG. 1: **Manifold minimization. (a)** When $d/w > 0.168$, traditional "soap-film" minimal surface that connects the two circle boundaries degenerates from a catenoid to two planar disks and destroys transportational functionality. Instead, keeping the length of every systole (shortest closed trajectory) fixed as w, the systolic minimal surface maintains transportational functionality. In general, every cylindrical surface as well as their combinations (by sewing their ends to form a tree network) is a systolic minimal surface. **(b)** Systolic surface minimizer: (Step 1) Choose n punctures on the Riemann sphere, where n is the number of terminals. (Step 2) Find the corresponding Jenkins–Strebel quadratic differential $q(z)dz^2$ [12] and calculate the horizontal (blue) and vertical (red) trajectories. Along both trajectories square-like quad meshes [13] are tiled over the manifold, giving rise to multiple charts (different colors) that are cylindrical surfaces (by letting each quad mesh have the same size). (Step 3) Fix n terminals accordingly in the 3D Euclidean space. Immerse the manifold into the Euclidean space isometrically. (Step 4) While keeping the immersion isometric, adjust l and τ of each cylindrical surface so that the overall surface area is minimized.

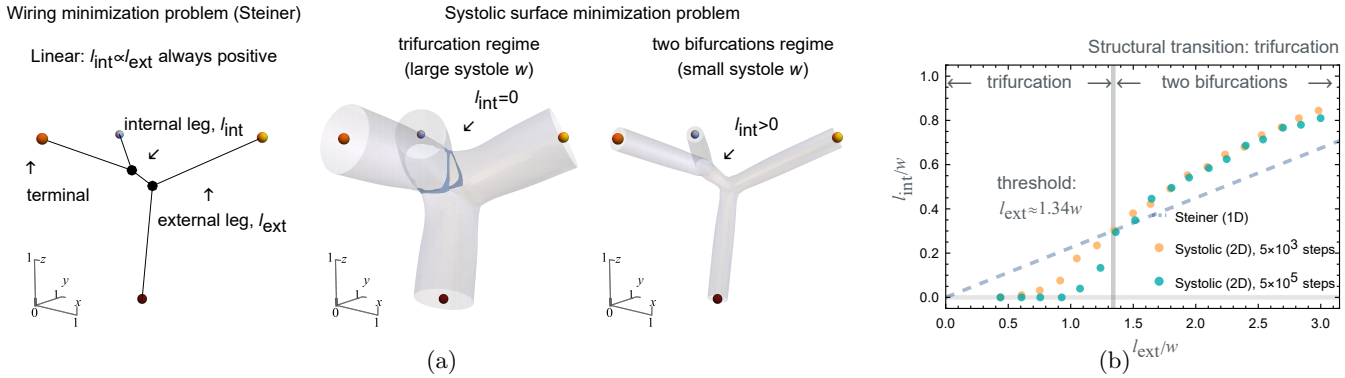


FIG. 2: **Trifurcation. (a)** Unlike Steiner graphs, systolic minimal surfaces allow trifurcation. **(b)** Given $n = 4$ terminals following perfect tetrahedral geometry, rather than linearly increasing with the internal leg length l_{int} (as in Steiner graph), the external leg length l_{ext} changes abruptly from the trifurcation regime ($l_{\text{int}} = 0$) to the double bifurcation regime ($l_{\text{int}} > 0$), crossing near the threshold of $l_{\text{int}} \approx 1.34w$.

tional functionality for finite w, its area always linearly scaling with its length. Although it is difficult to find a general numerical solution given general boundary conditions, it is proved that every *cylindrical surface*, which

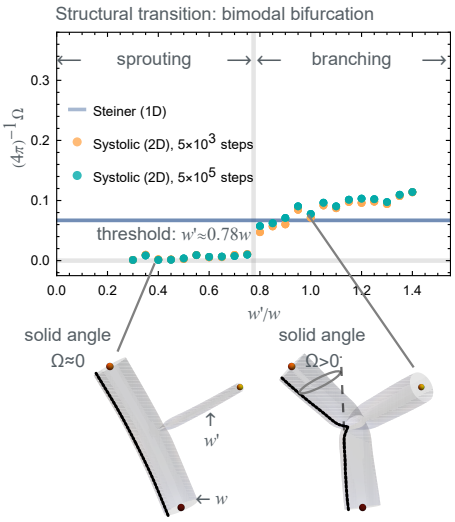


FIG. 3: **Bimodal bifurcation.** A bimodal bifurcation denotes a bifurcation with two typical but different systoles, one systole w for two of the external legs and another systole w' for the third external leg. Under manifold minimization, we observe a structural transition from the sprouting regime ($w'/w \lesssim 0.78$) to the branching regime ($w'/w \gtrsim 0.78$).

has a flat Riemannian metric everywhere (equivalent to a wrapped piece of flat paper but not necessarily a cylinder), is a *systolic minimal surface*, a special solution to our minimization problem [12]. Moreover, any 2D manifold, constructed from a tree graph of cylindrical surfaces as charts, is itself a systolic minimal surface too [15].

Hence, the systolic surface minimization problem reduces to a problem of first constructing a tree of cylindrical surfaces in the ambient space, i.e., an *isometric immersion* problem [13] that can be efficiently solved using the quad-mesh representation by letting each mesh not only be a square but also have the same size [13]. Given fixed boundaries, after finding all possible systolic minimal surfaces (i.e., with different twist τ or different l [Fig. 1(b)]) as solutions, the optimal solution is reached by selecting the systolic minimal surface with the smallest surface area from all possible combinations of l, τ, and node geometries.

Trifurcation.—Now we consider the systolic surface minimization problem with $N = 4$ terminals, located at the four corners of a perfect tetrahedron. Only l_{ext} and l_{int}, initially chosen to match the Steiner solution in the $w \to 0$ limit, are freely adjustable for surface minimization. When the systole $w > 0$, we find that l_{ext} and l_{int} follow a nonlinear relation that differs from Steiner's trivial linear relation (Fig. 2). When w is small (l_{ext}/w is large), the geodesic l_{int} of the internal leg remains positive and slightly above the Steiner prediction; when w is large (l_{ext}/w is small), however, l_{int} approaches zero. This indicates that a structural transition happens near a threshold of $l_{\text{ext}}/w \approx 1.34$, below which the degeneration of l_{int} signals the emergence of trifurcation.

Bimodal bifurcation.—A possible generalization of the systolic surface minimization is to constrain different physical links by different systoles. Here, we investigate the simplest case of a *bimodal* bifurcation that has different systoles, one systole w for two external legs and another w' for the third external leg, with all $N = 3$ terminals located at the three corners of an equilateral triangle (Fig. 3). Our algorithmic solution predicts not only that the branching angle (the solid angle Ω between the two same-systole external legs) correlates positively with the ratio of systoles w'/w, but also that a structural transition emerges near $w'/w \approx 0.78$, distinguishing two long-hypothesized different morphological regimes: [16] the "sprouting" regime (mode I [16]), where the branching angle is strictly zero; and the "branching" regime (mode II [16]), where the branching angle starts increasing with w'/w. Note also that bimodal bifurcations remain planar, hence do not violate the geometric constraint of planarity [9].

[1] N. Dehmamy, S. Milanlouei, and A.-L. Barabási, Nature **563**, 676 (2018).
[2] M. Barthélemy, Physics Reports **499**, 1 (2011).
[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[4] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Physical Review Letters **89**, 258702 (2002).
[5] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem*, 1st ed. (Elsevier, Amsterdam, The Netherlands, 1992).
[6] J. M. L. Budd, K. Kovács, A. S. Ferecskó, P. Buzás, U. T. Eysel, and Z. F. Kisvárday, PLOS Computational Biology **6**, e1000711 (2010).
[7] T. Latty, K. Ramsch, K. Ito, T. Nakagaki, D. J. T. Sumpter, M. Middendorf, and M. Beekman, Journal of The Royal Society Interface **8**, 1298 (2011).
[8] P. Bradley and M. Berry, Brain Research **109**, 133 (1976).
[9] Y. Kim, R. Sinclair, N. Chindapol, J. A. Kaandorp, and E. D. Schutter, PLOS Computational Biology **8**, e1002474 (2012).
[10] J. C. Chappell, D. M. Wiley, and V. L. Bautch, Seminars in Cell & Developmental Biology Mechanochemical Cell Biology, **22**, 1005 (2011).
[11] A. I. Bobenko, J. M. Sullivan, P. Schröder, and G. M. Ziegler, eds., *Discrete Differential Geometry*, 1st ed., Oberwolfach Seminars, Vol. 38 (Birkhäuser, Basel, Switzerland, 2008).
[12] K. Strebel, *Quadratic Differentials*, 1st ed. (Springer, Heidelberg, Germany, 1984).
[13] C. Jiang, C. Wang, F. Rist, J. Wallner, and H. Pottmann, ACM Transactions on Graphics **39**, 128:128:1 (2020).
[14] J. C. C. Nitsche, Journal of Mathematics and Mechanics **13**, 659 (1964).
[15] B. Zwiebach, Physics Letters B **241**, 343 (1990).
[16] M. Zamir, Journal of Theoretical Biology **62**, 227 (1976).

Analytic solution for the spectral density and localization properties of complex networks

Jeferson D Silva[1] and Fernando L Metz[1,2]

[1] Physics Institute, Federal University of Rio Grande do Sul, 91501-970 Porto Alegre, Brazil.

[2] London Mathematical Laboratory, 8 Margravine Gardens, London W6 8RH, United Kingdom.

Large complex systems are ubiquitous in the real world, ranging from physical and biological to social and technological systems [1]. The adjacency random matrix plays a central role in this context, as it describes the interactions among the individual elements that compound these large complex systems. The empirical spectral density of the adjacency random matrix and the localization of its eigenvectors are key quantities to understand a variety of dynamical processes on complex networks (or random graphs) [2, 3]. The spectral and localization properties of the adjacency random matrix are given by functions of the diagonal elements G_{ii} of the resolvent matrix $\mathbf{G}$. The imaginary part of G_{ii} determines the local density of states (LDOS), which counts the number of states at a certain eigenvalue λ at node i. The average of the LDOS over all the nodes determines the empirical spectral density, while The average of $|G_{ii}|^2$ gives information about eigenvector localization throughout the inverse participation ratio (IPR), which characterizes the volume of the eigenvectors. The probability density function of G_{ii} satisfies a system of distributional equations [4, 5, 6], providing a solid foundation to study the spectral and localization properties of *heterogeneous* random graphs. Heterogeneity is broadly associated with local fluctuations in the graph structure, such as randomness in the degrees or in the interaction strengths between the nodes (the degree of a given node counts the number of edges attached to it). Although the resolvent distributional equations have led to enormous progress, they admit analytical solutions only for random graphs with a homogeneous structure [5, 7, 8]. In a recent paper [9], the resolvent equations for the configuration model of random graphs with a geometric degree distribution have been studied in the high connectivity limit, i.e., when the average degree c becomes infinitely large. It is shown in this paper that the average resolvent satisfies a transcendental equation, and the spectral density diverges at the center of the spectrum. These findings are interesting because they suggest the existence of a new class of solutions for the distributional equations of the resolvent in the high connectivity regime, which lies between the sparse (when the average connectivity is finite) and the dense regime (when the random graph becomes fully connected). Moreover, these analytical results also imply that the spectral density of random graphs in the high connectivity limit is not typically governed by the Wigner semicircle law of random matrix theory [10], as it is rigorously proven in [11]. Indeed, the Wigner law universality only holds for random graphs with degree distributions that become highly concentrated around its mean value for $c \to \infty$. In other words, the average connectivity is large, but the fluctuations in the network are still relevant for the spectral properties. The analytical results obtained in [9] are limited, however, to a geometric degree distribution. In this work [12], we generalize the results of [9] and derive analytical solutions for the resolvent distributional equations of random graphs with arbitrary degree distributions in the high-connectivity limit. In this context, we perform a detailed analysis of the impact of degree fluctuations on the spectral density, the inverse participation ratio, and the distribution of the local density of states. For random graphs with a negative binomial degree distribution, we show that all eigenvectors are extended and that the spectral

density unveils a logarithmic or a power-law divergence when the variance of the degree distribution is sufficiently large. We elucidate this singular behaviour by showing that the distribution of the LDOS at the center of the spectrum exhibits a power-law tail determined by the variance of the degree distribution. In addition, we show that in the regime of weak degree fluctuations the spectral density of random graphs with a negative binomial degree distribution has finite support, which promotes the stability of large complex systems on random graphs.

We consider a simple and undirected random graph with N nodes. The network topology is specified by the components of the adjacency random matrix $\mathbf{A}$. We generate $\mathbf{A}$ according to the configuration model of networks [1, 13, 14] in which a random graph is chosen uniformly at random from the set of all random graphs with a given degree sequence $K_1, \ldots, K_N$ generated from a prescribed degree distribution p_k. In the high connectivity limit, the spectral density and the inverse participation ratio are, respectively, given by

$$\rho_\epsilon(\lambda) = \frac{1}{\pi}\mathrm{Im}\bigg[\int_0^\infty d\kappa \frac{\nu(\kappa)}{z - \kappa J_1^2\langle G\rangle}\bigg], \tag{1}$$

$$\mathcal{I}_\epsilon(\lambda) = \frac{\epsilon}{\pi\rho_\epsilon(\lambda)}\int_0^\infty d\kappa \frac{\nu(\kappa)}{|z - \kappa J_1^2\langle G\rangle|^2}, \tag{2}$$

with $z = \lambda - i\epsilon$ lying on the lower complex half-plane and J_1^2 denoting the variance of the distribution that defines the coupling strengths between the graph nodes. The variable $\langle G\rangle$ satisfies the fixed-point equation

$$\langle G\rangle = \int_0^\infty d\kappa \frac{\nu(\kappa)\,\kappa}{z - \kappa J_1^2\langle G\rangle}. \tag{3}$$

In the high connectivity limit, the fluctuations in the random graph are captured by the empirical distribution of the re-scaled degrees $\nu(\kappa)$, which is defined as

$$\nu(\kappa) = \lim_{c\to\infty}\sum_{k=0}^{\infty} p_k \delta\bigg(\kappa - \frac{k}{c}\bigg). \tag{4}$$

The solution of the self-consistent equation given by (3) determines each and every equation of this work.

For a negative binomial degree distribution $p_k^{(b)}$, one can investigate the role of degree fluctuations on the spectral and localization properties of random graphs in terms of a single parameter in the high connectivity limit, given by the relative variance of $p_k^{(b)}$, i.e.

$$\frac{1}{\alpha} = \lim_{c\to\infty}\frac{\sigma_b^2}{c^2}. \tag{5}$$

By considering the negative binomial distribution, we obtain a simple expression for the distribution of the LDOS at $z = 0$, viz.

$$P_0(y) = \frac{\alpha^\alpha}{\Gamma(\alpha)J_1^\alpha}\frac{e^{-\frac{\alpha}{J_1 y}}}{y^{\alpha+1}}. \tag{6}$$

Equation (6) reveals the unbounded character of the LDOS fluctuations at $\lambda = 0$. We show in this work that the spectral density diverges for $\alpha \leq 1$ at $\lambda = 0$ (see Figure 1). In this regime, the above result helps us to clarify this singular behaviour. The power-law tail of (6) exhibits a divergence in the q-th moment $\overline{y^q} = \int_0^\infty dy y^q P_0(y)$ for $\alpha \leq q$, which explains the singularity of the spectral density at $\lambda = 0$, for $\alpha \leq 1$. In addition, the non-singular, $\epsilon-$independent behaviour of (6), confirms the extended phase of the eigenvectors in the high connectivity limit at $\lambda = 0$.

In summary, this work unveils non-trivial results for the resolvent distributional equations of undirected random graphs with a heterogeneous structure in the high connectivity limit. All of our results are determined solely in terms of the empirical spectral density of the re-scaled degrees and the complex variable $\langle G\rangle$, in which the latter satisfies a self-consistent equation.

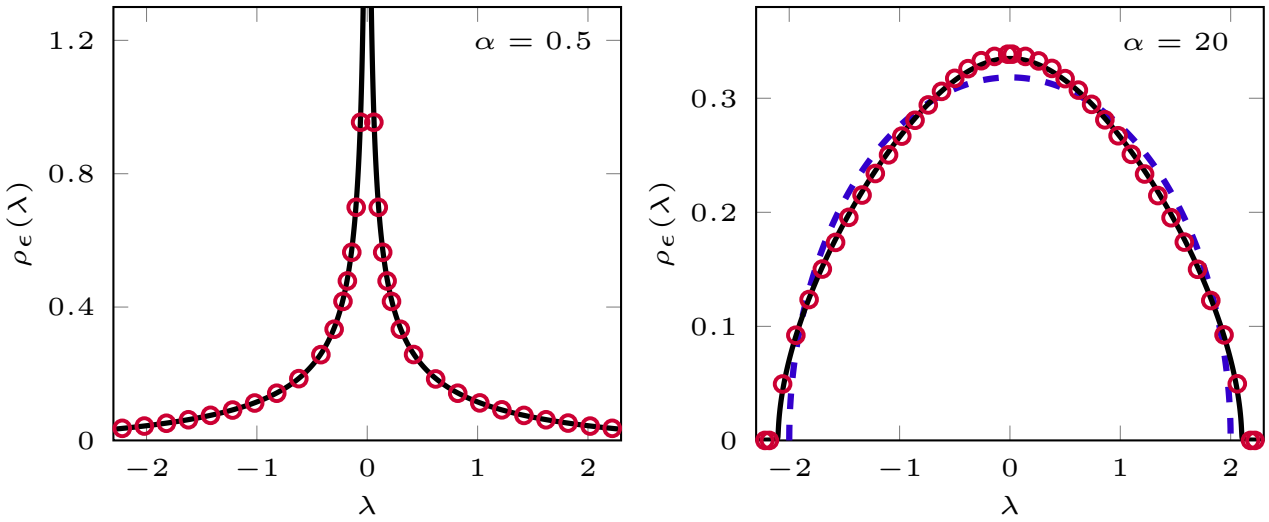


Figure 1: The spectral density of random graphs with a negative binomial degree distribution in the high-connectivity limit. The parameter $1/\alpha$ controls the relative variance of the degree distribution (see Eq. (5)). The solid lines are the theoretical results derived from solving Eqs. (1) and (3) for $\epsilon = 10^{-3}$ and $J_1 = 1$. The red circles are numerical diagonalization results obtained from an ensemble of $10^4 \times 10^4$ adjacency random matrices. The dashed blue curve in the right panel represents the Wigner semicircle law.

References

[1] Newman M E J, Strogatz S H and Watts D J 2001 *Phys. Rev. E* **64**(2) 026118

[2] Restrepo J G, Ott E and Hunt B R 2006 *Phys. Rev. Lett.* **97**(9) 094102 URL `https://link.aps.org/doi/10.1103/PhysRevLett.97.094102`

[3] Martin T, Zhang X and Newman M E J 2014 *Phys. Rev. E* **90**(5) 052808 URL `https://link.aps.org/doi/10.1103/PhysRevE.90.052808`

[4] Dean D S 2002 *Journal of Physics A: Mathematical and General* **35** L153

[5] Rogers T, Castillo I P, Kühn R and Takeda K 2008 *Physical Review E* **78** 031116

[6] Kühn R 2008 *Journal of Physics A: Mathematical and Theoretical* **41** 295002

[7] Bordenave C and Lelarge M 2010 *Random Structures & Algorithms* **37** 332–352

[8] Metz F L, Neri I and Bollé D 2011 *Physical Review E* **84** 055101

[9] Metz F L and Silva J D 2020 *Physical Review Research* **2** 043116

[10] Livan G, Novaes M and Vivo P 2018 *Monograph Award* 63

[11] Dembo A, Lubetzky E and Zhang Y 2021 Empirical spectral distributions of sparse random graphs *In and Out of Equilibrium 3: Celebrating Vladas Sidoravicius* (Springer) pp 319–345

[12] Silva J D and Metz F L 2022 Analytic solution of the resolvent equations for heterogeneous random graphs: spectral and localization properties URL `https://arxiv.org/abs/2209.06805`

[13] Molloy M and Reed B 1995 *Random Structures & Algorithms* **6** 161–180

[14] Fosdick B K, Larremore D B, Nishimura J and Ugander J 2018 *SIAM Review* **60** 315–355

Assessing the effectiveness of perimeter lockdowns as a response to epidemics at the urban scale: the case of Madrid

Alfonso de Miguel Arribas[1,2], Alberto Aleta[1,2] and Yamir Moreno[1,2,3]

[1] *Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018, Zaragoza, Spain.*

[2] *Department of Theoretical Physics, University of Zaragoza, 50018, Zaragoza, Spain. and*

[3] *Centai, Torino, Italy.*

(Dated: November 7, 2022)

During the COVID-19 pandemic we have become acquainted with a battery of measures to fight the spreading of the infection produced by SARS-CoV-2 at different scales. At first, there was little option other than pursue non-pharmaceutical interventions (NPIs) as travel bans and lockdowns of different extension. At the urban level, one of these tools has been the so-called localized or perimeter lockdowns. When the epidemiological situation of an area worsens, perimeter lockdowns aim for protection to the rest of the system by banning travels in and out, and reinforcing awareness and action on the affected areas. This strategy is argued to result in lower social and economic costs as compared to larger-scale restrictions, but there are also some concerns about their usefulness in effectively controlling an epidemic at a urban scale. This strategy is a rare tool only implemented in Santiago de Chile and Madrid to our best knowledge [1]. The case of Madrid caused certain sociopolitical controversy, and even jumped into scientific literature [2], [3]. Inspired by this, in this work we try to settle the question on the effectiveness of perimeter lockdowns (PLs).

We use a data-driven stochastic metapopulation SIR epidemiological model of a city which responds to the epidemic spreading with PLs. Our model thus consists in a population of N individuals distributed in a networked structure of V patches that represent urban districts. Patches are connected through data-driven mobility flows. Inside every subpopulation the homogeneous mixing assumption is considered. Apart from the disease parameters, related in this settings through $R_0 = \beta T_I$, where R_0 is the basic reproductive number, β is the transmissibility rate and T_I is the infectious period, there is also a general mobility parameter κ, a transmissibiity reduction parameter χ, and a risk incidence threshold Θ. When a subpopulation in the system shows a 14-day cumulative incidence rate above a certain threshold Θ, travel restrictions in and out of the affected area i are activated and a local transmissibility reduction is carried through lowering χ, so that $\beta_i = \chi_i \beta$.

We explore under which circumstances PLs could be a good response of epidemic control and we find that the window of opportunity is very tight, making them rather useless in most realistic situations. Mobility reductions by itself do nothing unless κ is moved to unrealistically low values. Indeed, achieving high enough transmissibility reductions, that is, low χ, is key to locally control the spreading but even more importantly is to act as soon as possible, very low Θ and this could not be that easy to achieve. Given the interconnectedness of the system, synchronization of the epidemic trajectories in every subpopulation takes place and the full system is quickly invaded and at risk. This synchronization is something

that can be seen by simple inspection of the real data in Madrid. Our parsimonious model reproduces these qualitative aspects well. This phenomenon, due to highly interconnected mobility flows among districts, is what hinders the effectiveness of PLs at the urban scale.

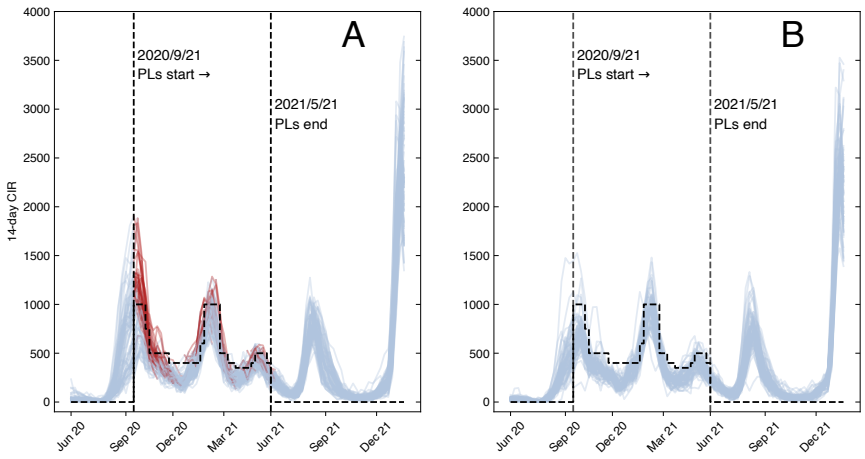


FIG. 1: Real 14 day cumulative incidence rate time series for the basic health zones (BHZs) in Madrid city. Left: Trajectories for BHZs that during some time period experienced a PL. In red, the period in which they were under a PL. Right: Trajectories for BHZs that never were confined. Vertical dashed lines mark the beginning and the end of the perimeter lockdown strategy in Madrid and step-wise horizontal dashed lines signal the risk threshold considered by the authorities to activate the PLs.

[1] Li Y, Undurraga EA, Zubizarreta JR. Effectiveness of Localized Lockdowns in the COVID-19 Pandemic. American journal of epidemiology. 2022.

[2] Fontán-Vela M, Gullón P, Padilla-Bernáldez J. Selective perimeter lockdowns in Madrid: a way to bend the COVID-19 curve? European Journal of Public Health. 2021;31(5):1102-4.

[3] David GG, Rafael HH, Ayelén RB, Inmaculada LG, Amparo L, Marina P, et al. Perimeter Confinements of Basic Health Zones and COVID-19 Incidence in Madrid, Spain. BMC Public Health. 2021;22:216.

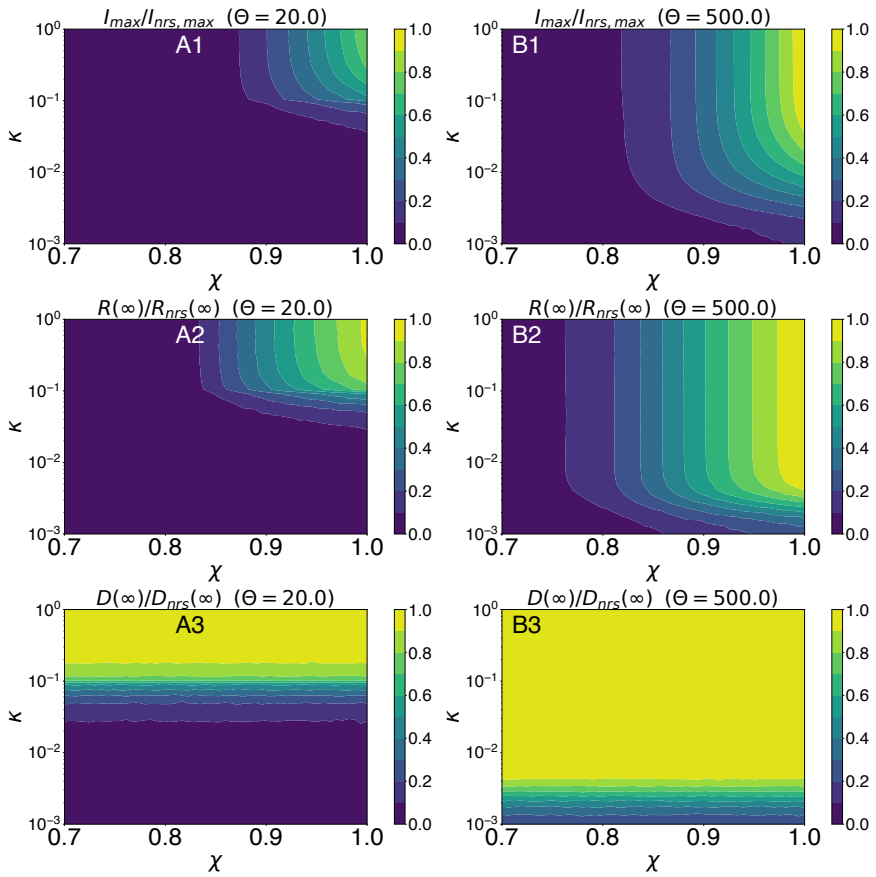


FIG. 2: Color maps show epidemic impact when varying simultaneously κ and χ. Plots A1, A2, and A3 show the peak incidence, prevalence, and locked districts fraction, respectively, in (χ, κ)-space for threshold $\Theta = 20$. Plots B1, B2 and B3 show the same observables now for $\Theta = 500$. Quantity values are normalized with respect to a no-response scenario.

Attraction by ingroup coherence drives the emergence of ideological sorting

Lucía Pedraza[1], Federico Zimmerman[2], Joaquín Navajas[2], and Pablo Balenzuela[1]

[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IFIBA (Instituto de Física de Buenos Aires) - CONICET.
[2]Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina.

Abstract

In this work, we propose an agent-based model to study the mechanism underlying polarization and ideological sorting in a group of agents. Based on experimental results, we assume that the two main mechanisms driving interactions are homophily, in which people who share similar opinions are more likely to interact, and in-group-coherence favoritism, in which they are more attracted to coherent in-groups rather than incoherent or out-group members. Interestingly, by incorporating this latter assumption, we were able to observe ideological alignment as observed in political opinions. Additionally, by solving the model's master equations and fitting the model's parameters with opinions from survey's data and experiments we show that a combination of homophily and in-group coherence is the best explanation for the emergence of ideological sorting.

Keywords— opinion model, ideological sorting, agent based models, master equations

1 Introduction

The increasing political polarization has become a worrying concern in many different countries and a serious threat to society and democracy itself. In response to this scenario, interdisciplinary efforts combining empirical and theoretical approaches, have sought to better understand the involved mechanisms and the conditions under which societies are expected to polarize.

One promising way to address this question is by studying the behavior of agent-based models (ABMs) under different conditions of social influence and interactions. Based on a few simplistic assumptions, most studies have focused on under which conditions individuals may reach consensus or polarize and become more extreme across one single topic, presumably oversimplifying the actual complexity of the current landscape and the underlying social dynamics.

While disagreement on policy issues is increasing, opinions are also becoming more extreme but also more coherent across diverse and seemingly unrelated topics. In several countries people have ideologically sorted in terms of partisanship leading to an issue alignment. For example, in the US, an individual who supports the women's right to voluntarily terminate pregnancy will be more likely to support stricter legislation on gun control, even though these topics are, a priori, unrelated to each other.

A recent paper [1] have shown that people not only hold politically coherent opinions across very different issues but also that this property, i.e., political coherence, increases interpersonal attraction among co-partisans. In other words, individuals who hold coherent opinions are more attractive than those individuals having some degree of ambivalence in their attitudes (e.g., a person who is anti-abortion but supports gun control). However, whether and how this driver of interpersonal attraction relates to macro-level patterns of political polarization and partisan-ideological sorting remains largely unknown.

Interestingly, by incorporating this latter novel experimental finding from political psychology at the microscopic level, issue alignment emerges starting from a random distribution of opinions. We formulated the model, derived the master equation of the system and performed numerical simulations obtaining identical results. Additionally, we compared different final states of the model with data from multiple datasets (which include more than 20,000 opinions on different controversial issues). By fitting the model's parameters with the data, we were able to observe that homophily alone could not explain ideologically sorted states and enhancing the importance of considering ingroup-coherence favoritism in political interactions.

2 The model

We consider a system of N agents. Each one holds a multidimensional opinion, where each dimension reflects the agent's opinion on a particular issue. For the sake of simplicity, we will consider a debate on only two issues. Agent i will hold an opinion $O_i = (O_i^1, O_i^2)$ where each coordinate could be against (-1), neutral (0) or in favour (1). For convenience, when possible, we assume that every political issue expresses a right-wing opinion facilitating us to notate right-wing opinions as positive, left-wing opinions as negatives, and undecided as zero. We also classify all agents by their ideological consistency. This procedure leads to following different communities:

- Coherent agents (C): hold assertive and matching opinions on both issues ($O = (-1, -1)$ or $O = (1, 1)$).
- Incoherent agents (I): hold assertive but opposite opinions on the two different issues ($O = (-1, 1)$ or $O = (1, -1)$).
- Weak agents (W): Agents that hold an assertive opinion in one issue and are undecided regarding the other one ($O = (-1, 0)$, $O = (0, -1)$, $O = (1, 0)$ or $O = (0, 1)$).
- Apathetic agents (A): Agents that are undecided on both issues ($O = (0, 0)$).

Over time, agents interact with each other and these interactions influence agents' opinions. At each time step, two agents are randomly selected and their interaction would lead one of the agents to influence the other with probability P. We incorporate two different mechanisms that impact agents' influence coupled by a parameter k.

- **Homophily** We define similarity as one minus the normalized Manhattan distance between the agents' opinions (L_1 distance, i.e. the absolute value of the sum of the distance in both coordinates): $S_{ij} = 1 - \frac{|O_i - O_j|_1}{4}$
- **Ingroup coherence favoritism** We define agents' ideology as the sum of the opinions on the two issues and we normalize it by dividing it by two: $I_i = \frac{O_i^1 + O_i^2}{2}$. The ideology sign value corresponds to the agent's leaning, a positive value corresponds to right-wing agents and negative one to left-wing agents. The ingroup coherence displayed by agent i to agent j is the absolute value of the ideology of i, if both are in the same group ($\sigma_{ij} = 1$), and zero otherwise: $C_{ij} = |I_i|\sigma_{ij}$.

The probability of influence is defined $P_{ij} = (1 - k)S_{ij} + kC_{ij}$
For high values of k only similarity is taking action and for low values only coherence.

Pairwise interactions lead to opinion changes that could occur only in one of the two issues and only by one unit. When agents are similar, influence will be attractive and agent i will move closer to j by changing one of its own opinions. Conversely, for dissimilar agents interactions are repulsive. In this case, agent i will change one of its opinions moving further from agent j and reducing their similarity.

3 Simulations, analytic results and empirical data

To derive the master equation, we consider that at each interaction, two agents are randomly selected. Therefore, the probability that these agents belong to a community is the fraction of agents in the communities. Then, we compute the flux between populations by considering the likelihood of changing an opinion after an interaction, obtaining the following equations:

$$\begin{aligned}
\frac{dC}{dt} &= WC(\frac{k}{16} + \frac{1-k}{4}) + W^2\frac{k}{16} \\
\frac{dI}{dt} &= WI\frac{k}{16} + W^2\frac{k}{16} \\
\frac{dA}{dt} &= -AC\frac{k}{2} - AI\frac{k}{2} \\
\frac{dW}{dt} &= -\frac{dC}{dt} - \frac{dI}{dt} - \frac{dA}{dt}
\end{aligned}$$

These equations were numerically solved with randomly distributed initial conditions such that: $C(0) = I(0) = 2/9$, $A(0) = 1/9$ y $W(0) = 4/9$. For numerical simulations of the ABM , we consider a system with N=1000 agents and averaged the results of 100 simulations per condition.

In Figure 1 (A) we show the proportion of coherent and incoherent agents, depending on the parameter k. The parameter k was varied from 0 to 1 with steps of 0.05. We can see the perfcet match between simulations and the theoretical results. In all cases (except $k = 1$), only the coherence and incoherence agents remain whereas the apathetic and the weaks disappear. For higher k, the coherence mechanism is stronger leading the system to a final state where the coherence community dominates the population.

The model's simulations and the analytical formulation display a clear relationship between the influence of in-group coherence favoritism and the final proportion of coherent agents. Next, we analyze to what extent actual opinions on a great variety of controversial issues are sorted. We work with multiple datasets [2, 3, 4] with more than 20,000 responses on different polarizing topics. All responses indicate the participants' agreement,

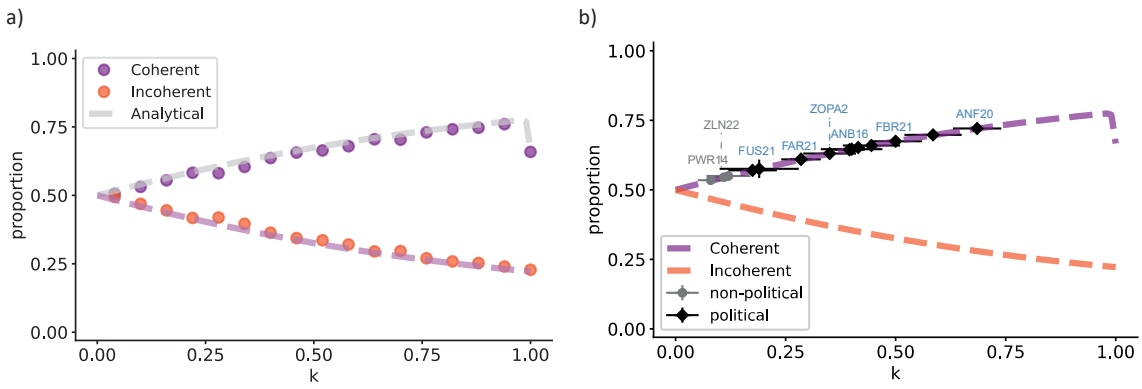


Figure 1: A) The model's final states are shown for different values of k. The figure depicts the mean values of 10 different simulations per scenario. The final proportion of coherent agents is shown in purple and the final proportion of incoherent agents in orange. For $k < 1$, as k increases, so does the final proportion of coherent agentes. B) Every dataset's mean sorting value was mapped to its corresponding model's k value. Non-political datasets are shown in light gray and political ones in black.

disagreement or neutrality to each different issue. In order to contrast the data to the proposed two-dimensional model, we focused on the relation between the proportion of coherent (C) and incoherent (I) agents and did not consider in the analysis the weak and apathetic populations. Thereby, we compute sorting ($S = \frac{C}{C+I}$) for all the possible pairs of opinions within each dataset.

Having the model's analytical solution allows us to map every sorting value into its corresponding model's k parameter (Figure 1(B)). Non-political datasets are shown in light gray dots and black diamonds are used for political datasets. We note that the value of k (the proportion of ingroup coherence) divide non-political from political groups of opinions. Taken altogether, data suggests that homophily alone can not explain the emergence of the observed levels of sorting and political polarization. On the contrary, non-political controversial opinions exhibit the lowest observed levels of sorting which can be explained by homophily and attractive-repulsive interactions.

References

[1] Federico Zimmerman et al. "Political coherence and certainty as drivers of interpersonal liking over and above similarity". In: *Science advances* 8.6 (2022), eabk1909.

[2] *The American National Election Studies (ANES)*. https://electionstudies.org/data-center.

[3] *Pew Research*. https://www.pewresearch.org/.

[4] Lucia Freira et al. "The interplay between partisanship, forecasted COVID-19 deaths, and support for preventive policies". In: *Humanities and Social Sciences Communications* 8.1 (2021), pp. 1–10.

Authority without Care: Moral Values behind the Mask Mandate Response

Yelena Mejova[1], **Kyriaki Kalimeri**[1], **Gianmarco De Francisci Morales**[2]
[1]ISI Foundation, Turin, Italy; [2]Centai, Turin, Italy
Keywords: *Twitter, mask wearing, moral values, topic modeling*

Face masks are one of the cheapest and most effective non-pharmaceutical interventions available against airborne diseases such as COVID-19. In the U.S., masks have been met with resistance by a substantial fraction of the populace. Being a prosocial behavior, mask-wearing is influenced by our political ideology [1] and moral values [2, 3] which are directly linked to moral decision making [4]. In this work, we provide a fine-grained analysis of the moral values of those expressing opinions around masking in the U.S. by applying the Moral Foundations Theory to a dataset of Twitter posts spanning the beginning of the pandemic, from January to July 2020. In particular, we ask: *What is the anatomy of the collective discussion on mask wearing around the mask mandate on Twitter?* In particular, focusing on the U.S., we analyze different facets of this discussion:

1. **RQ1.** How does their stance relate to their political leaning;
2. **RQ2.** What moral values do adherents to pro- or anti-masking stances hold;
3. **RQ3.** What is the information environment around their arguments?

Data Collection. We begin by collecting tweets mentioning the keywords "mask", "facemask", "ffp3", and "n95" (the latter two refer to popular kinds of masks), spanning the dates of January 1st to July 30th, 2020, using the GOT3 library [5]. These keywords were chosen by considering the special Twitter Covid-19, stream[1] and picking the most common English keywords related to masks. This collection results in $18\,245\,298$ tweets from $5\,935\,103$ users. After performing basic pre-processing steps, we employed $430\,568$ tweets to train a relevance classifier, maintaining only the tweets that were related to the pandemic. We geolocated the relevant tweets by direct string matching of the declared location of the user in the Location strings to the Geonames ID. Finally, we used the Twitter API Friends call to collect the information about whom these users follow ("followees" or "friends"), thus resulting in the coverage of $598\,792$ users.

Stance Classification. Following previous work on identifying controversial topics on social media [6, 7], we look for a bi-partitioning of the network that would indicate polarization. Constraining the followees to the set of users in our tweet dataset, the network contains $|V| = 598\,792$ users and $|E| = 35\,763\,336$ edges. We use the graph partitioning algorithm METIS [8] to partition the network into two groups,

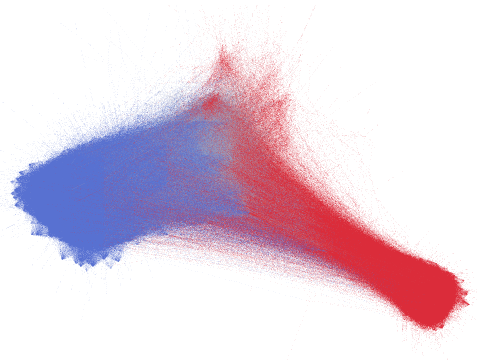


Fig. 1. GCC of follower network, colored by METIS score.

allowing us to assign a label to 56.4% of users: 28.8% with 0 and 27.6% with 1, thus leaving 43.5% of users with an unknown label. The two sides are colored as blue and red, and unknown as grey in Figure 1. Manual annotation of a sample of users from each stance revealed an overall precision of 86.4%, with perfect precision for pro-mask class, but only 72.4% for anti-mask case, with several users, incorrectly labeled as anti-mask by the algorithm. The network structure suggests some connection of people who express doubts, although are not clearly anti-mask, with more extreme positions.

Political Leaning. From the mask stance of the users we are able to classify to their political affiliation, which can be glimpsed via their Twitter social network. Previous literature suggests that users mostly follow accounts that are in agreement with their political views [9]. We created a list of prominent political accounts in order to propagate their leaning to their followers[2] includes 501 accounts of members of the U.S. Congress, 79 governors, 70 party entities, and 67 Attorney Generals, as well as 157 media accounts from allsides.com[3] and 67 journalists from politico.com[4].

We consider only users who follow at least 5 accounts in our list from either side, and calculate the aggregated political leaning score as $S_{PL} = (N_R - N_L)/(N_R + N_L)$, which results in $S_{PL} \in [-1, 1]$ with 1 the most right-leaning score. Thus, we are able to identify the political leaning of $18\,422$ users. Pro-mask users are more likely to be following left-leaning accounts, and anti-mask ones the right-leaning ones, with almost no users existing in the middle political ground.

[1]https://developer.twitter.com/en/docs/twitter-api/tweets/covid-19-stream/filtering-rules

[2]Available at https://tinyurl.com/poliaccounts

[3]https://www.allsides.com/media-bias/media-bias-ratings

[4]https://www.politico.com/blogs/media/2015/04/twitters-most-influential-political-journalists-205510

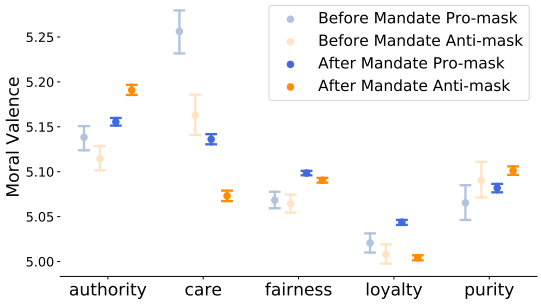


Fig. 2. Moral valence in narratives expressed by pro-mask and anti-mask users in the periods before (lighter points) and after (darker points) the mandate. Dot represents the median value while the whiskers represent 5-95% quantiles.

Moral Values. Applying the MoralStrength lexicon [10] on all tweets we obtained an average moral score per foundation for each tweet. Our results show that, while the anti-mask stance is associated with a conservative political leaning, the moral values expressed by its adherents diverge from the ones typically used by conservatives. Figure 2 shows the mean moral value scores of each side in the periods before and after the mandate. Before, the two sides display comparably similar values, except for *care*, which is by far higher for the pro-mask side (significant at $p < 0.001$). There is a clear shift in the moral narratives expressed after the mandate by both sides of the debate. First, we find an increase in the valence of *authority* for the anti-mask side ($p < 0.001$), which is mostly accompanied by criticism and mistrust of the decisions made by the authorities. The pro-mask side sees a lack of leadership in former President Trump's refusal to wear a mask. The fact that post-mandate the authority-related keywords have higher valence on the anti-mask side suggests stronger criticism of the authorities than the pro-mask side (for whom the increase is significant only at $p = 0.004$ before the Benjamini-Hochberg correction for multiple hypothesis testing).

In terms of *care*, both sides have a downwards shift after the mandate. For the pro-mask side, this shift is accompanied by an increase in *fairness* and *loyalty*, which can be interpreted as a shift in focus from personal choice based on caring for others to complying with the mandate. Conversely, anti-mask supporters express themselves by prioritizing much less the notion of *care*, explicitly showing disregard for the protection of others, or simply stating that they do not care about being criticized for not wearing a mask. In addition, pro-mask supporters express significantly more *loyalty* in their messaging after the mandate ($p < 0.0001$). After the mandate, the *fairness* value increases for both sides (both at $p < 0.0001$). The expected emphasis on the values of authority and purity is accompanied by an atypical dearth of in-group loyalty.

Complementing our previous analysis, the interrupted time series model shows that for all the moral dimensions, after the mandate, there is an evident change in behavior by both sides. The most interesting moral dimension is *loyalty*, whose signal is evidently diverging for two sides exactly after the mandate date and continues the same trend until the end of our data collection. We also observe that not

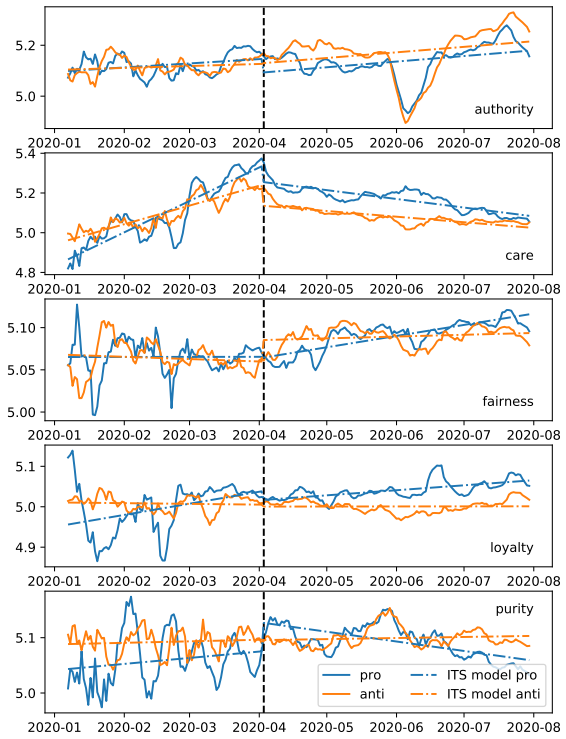


Fig. 3. Time series of moral value scores of pro-mask and anti-mask users, along with an interrupted time series analysis model.

Table 1. Use of singular and plural personal pronouns in a tweet by side, before and after the mask mandate.

Pronoun	Singular		Plural	
Mandate	before	after	before	after
Pro-mask	0.55	0.48	0.11	0.10
Anti-mask	0.53	0.60	0.09	0.10

only does the value of *care* decreases, the trend is downward over time, signaling a progressive shift in the debate. Similarly, the value of *purity* has a progressively negative trend for pro-mask side over time. Thus, we find that the temporal dimension of the data can be instructive about the evolution of the rhetoric in terms of divergence between the two sides of conversation and changes in emphasis.

Collectivism vs Individualism. One of the main purposes of mask wearing is the protection of others, an expression of solidarity within the in-group against an external threat. Thus, we turn to the Individualism-Collectivism (IC) dimension [11], which captures the standing of individuals as interdependent members of a collective. We operationalize it via the personal pronouns used in the tweets, mainly first-person singular ("I", "me","mine" etc.) and first-person plural ("we", "us", "ours" etc), following existing literature [12]. Table 1 shows that although having comparatively similar usages of singular pronouns before the mandate, the debate after the government's messaging becomes more individualistic for anti-mask side and less so for advocates of masking. The mask mandate reverses the expression of Individualism-Collectivism between the two sides, with an increase of individualism in the anti-mask narrative, and a decrease in the pro-mask one.

Table 2. Counts of the top 30 URL domains posted by pro- and anti-mask users. Domains colored by class: news and news aggregators (black), social media and social media automator/aggregators (red), business platforms (blue), medical organization (green).

Pro-mask		Anti-mask	
rawstory.com	2317	youtube.com	3485
cnn.com	2000	thegatewaypundit.com	1341
youtube.com	1751	etsy.me	1210
washingtonpost.com	1393	instagram.com	912
a.msn.com	935	foxnews.com	903
apple.news	872	zazzle.com	796
huffpost.com	758	breitbart.com	781
news.yahoo.com	686	nypost.com	472
flip.it	630	fineartamerica.com	453
nytimes.com	587	fxn.ws	393
nbcnews.com	573	westernjournal.com	362
thehill.com	527	dlvr.it	344
dailykos.com	486	pixels.com	317
instagram.com	458	buff.ly	288
businessinsider.com	449	theblaze.com	282
thedailybeast.com	402	bizpacreview.com	268
newsweek.com	371	infowars.com	250
theguardian.com	343	etsy.com	246
usatoday.com	337	ift.tt	236
yahoo.com	333	cnn.com	231
cnbc.com	307	twitchy.com	217
politico.com	301	ebay.us	202
newsbreakapp.com	295	ncbi.nlm.nih.gov	196
buff.ly	260	facebook.com	192
npr.org	252	nejm.org	191
politicususa.com	236	newsbreakapp.com	175
apnews.com	231	a.msn.com	173
nypost.com	223	google.com	162
latimes.com	209	dennismichaellynch.com	152
mol.im	208	aapsonline.org	142

Information Environment. Finally, we apply LDA to the argumentation obtained by each side to find major topics mentioned by either side. The most prominent one on the pro-mask side concerns the various interventions, including *social distancing* and *wearing* a mask. On the anti-mask side, the most prominent topic also concerns the interventions, but instead focuses on whether interventions *work* against the *spread*.

Finally, we turn to the information sources used by the two sides. Table 2 shows the top domains of the URLs posted by pro- and anti-mask users, along with the counts. Pro-mask users overwhelmingly post URLs pointing to news websites or aggregators. YouTube and Instagram feature prominently in both lists, though anti-mask users favor YouTube more than twice the second most popular domain. Anti-mask tweets also link to a variety of business platforms, including Etsy and Ebay, and lesser-known ones such as Zazzle, a platform for custom-designed products.

In conclusion, we note the lack of loyalty among the values emphasized by the anti-mask side, which tends to hold a conservative political view, and differs from the commonly observed ones associated with conservatism: authority, loyalty, and purity. This interpretation may point to motivated reasoning, wherein the desired conclusion modifies the worldview usually taken. Our findings suggest that there is an active development of symbolism and aesthetics of the resistance movement. Awareness of such symbolism and self-conceptualization is vital for crafting appropriate messages and fostering communication between the two sides. We argue that monitoring the dynamics of moral positioning is crucial for designing effective public health campaigns that are sensitive to the underlying values of the target audience.

[1] Jay J Van Bavel and Andrea Pereira. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3):213–224, 2018.

[2] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029, 2009.

[3] Peter K Hatemi, Charles Crabtree, and Kevin B Smith. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806, 2019.

[4] Sampada Karandikar, Hansika Kapoor, Sharlene Fernandes, and Peter K Jonason. Predicting moral decision-making with dark personalities and moral values. *Personality and Individual Differences*, 140:70–75, 2019.

[5] Dmitry Mottl. Getoldtweets3. https://pypi.org/project/GetOldTweets3/, 2021. (Accessed on April 1, 2021).

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying Controversy in Social Media. In *WSDM*, pages 33–42, 2016.

[7] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, 1(1):3, 2018.

[8] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392, 1998.

[9] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *SIGCHI conference on human factors in computing systems*, pages 1105–1108, 2011.

[10] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184, 2020.

[11] Geert Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[12] Jean M Twenge, W Keith Campbell, and Brittany Gentile. Changes in pronoun use in american books and the rise of individualism, 1960-2008. *Journal of cross-cultural psychology*, 44(3):406–415, 2013.

Brain drain or brain circulation? a study of the academic collaboration network in Uruguay

Pablo Galaso[1], Sergio Palomeque[2]

[1] Associate profesor in Economics
Universidad de la República - Instituto de Economía, Facultad de Ciencias Económicas y de la Administración, Gonzalo Ramírez 126, Montevideo 11200 (Uruguay)
+598 24131007
pablo.galaso@fcea.edu.uy

[2] Lecturer in Economics
Universidad de la República - Instituto de Economía, Facultad de Ciencias Económicas y de la Administración, Gonzalo Ramírez 126, Montevideo 11200 (Uruguay)
+598 24131007
sergio.palomeque@fcea.edu.uy

Draft version prepared for the NetSci-X 2023 Conference - November, 2022

Keywords: brain drain, brain circulation, coauthorship networks, ERGM, Latin America.

Abstract:

The emigration of highly qualified people is a challenge for the development processes of sending countries, particularly in cases such as Uruguay, which has been characterised for several decades as a country that has been an expeller of its qualified population. The literature analysing international migratory movements has broadened the view of "brain drain", which considers the negative effects on sending countries, by also exploring the channels of "brain circulation" and "brain gain", which can be generated through collaboration networks or the return of emigrants to their countries of origin.

This article seeks to make a contribution to the subject from the perspective of networks. In particular, the aim of the article is to analyse how the emigration and return of researchers can determine the formation of academic collaboration networks in Uruguay. As a general hypothesis, we propose that migration trajectories determine the configuration of researcher networks, resulting in a gain for the country from greater cooperation between those who reside abroad and those who reside in Uruguay, while returnees maintain relevant connections with other countries. From a network perspective, this hypothesis implies that both emigrant scientists and returnees tend to occupy gatekeeper positions, connecting local actors with agents abroad. We also expect this gatekeeper position to be determinant in the formation of the collaborative network between academics in Uruguay.

To test these hypotheses, we reconstructed the network of collaboration between academics to produce academic articles. In particular, we use the SCOPUS database, where we identify the authors linked to Uruguay using data from the First Census of People with PhD Degrees in Uruguay and all the public CVs of the Uruguayan national research and innovation agency (ANII). This allows us to extract, from the SCOPUS database, all papers with at least one author linked to Uruguay.

In order to analyse the emigration and return of Uruguayan academics, we used the institutional affiliation data reported in the authorship of each article. This allows us to differentiate between two broad categories: national affiliation (a university or research

centre located in Uruguay) and international affiliation (an institution abroad). From this distinction, we can identify, for each moment in time, those who emigrated, those who returned and those who never left the country. Likewise, the data allow us to include in the analysis those academics who always resided abroad, but who maintained some co-authorship link with academics in Uruguay.

To analyse the collaboration network we consider researchers as nodes and co-authorships as the (non-directed) links. The data includes papers between 1973 and 2021, which allows us to analyse the evolution of the network. To study the propensity of emigrants and returnees to occupy gatekeeper positions, we employ the categories used by Gould and Fernández (1989) as an indicator of this role in the network. In particular, we are interested in studying whether these actors mediate between local and foreign academics, as well as between other emigrants and/or returnees.

The analysis of the network topology shows a significant growth in the number of nodes together with an improvement in the cohesion of the network. Of particular interest is the emergence of a giant component, around the beginning of the 2000s, where about 60% of the network's nodes are connected (Figure 1).

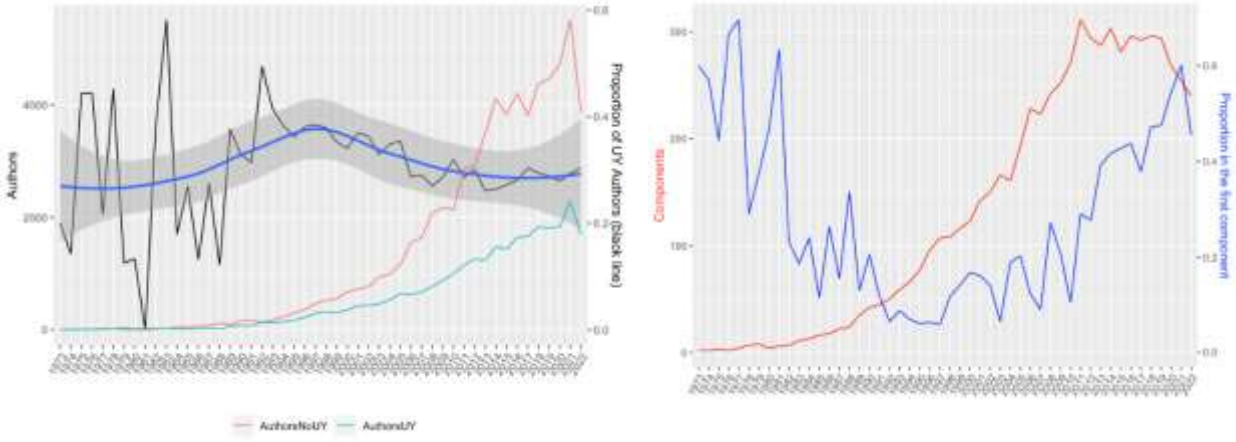


Figure 1. Number of author-nodes (left) and number of components vs. proportion of nodes connected to the largest component (right).

When we focus on the gatekeeper role of emigrants and returnees, the data reveals the propensity of movers (i.e. those who emigrate and those who return after a period abroad) to occupy this position. As shown in Figure 3, these researchers have clearly higher levels of intermediation than the rest of the researchers in the academic networks who are not movers (both national and foreign researchers).

Subsequently, in order to study the determinants of the network, we estimate models of network formation processes, in particular the so-called Exponential Random Graph Models (ERGM) (Robins et al. 2007). In these models we estimate an exponential random process of network formation that maximises the probability of the emergence of a network equal to the observed one. To account for the temporal evolution of the network, time-sliced ERGMs will be used (Kolaczyk and Csárdi 2014), which allow for separate modelling of link length distributions and structural dynamics of link formation.

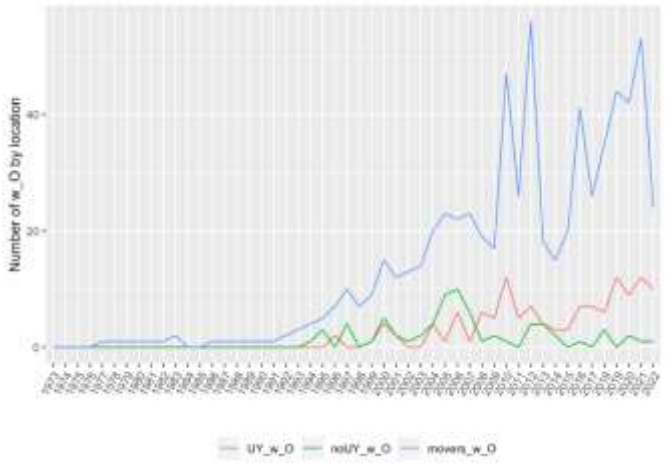


Figure 3. Proportion of researchers of each type occupying an intermediary position.

Note: the intermediary position involves connecting two unconnected researchers to each other, one of them being resident in Uruguay and the other non-resident.

Among the network configurations we estimate in the models, we focus on whether individuals are migrants or returnees. Thus, the ERGM models allow us to estimate whether emigration and return are determinants in the formation of the networks observed in our data. As control variables, we include other characteristics of collaborative structures for research and innovation that have been documented in previous empirical studies (Tomassello et al. 2017, König et al., 2011), such as geographical proximity, team formation or the presence of highly connected actors.

The results show an important role of emigrants and returnees in the shaping of the network. These types of actors seem to have a propensity to occupy gatekeeper positions. On the other hand, the first estimations made with the ERGM models indicate that migration and return are determinant in explaining the structure of the collaboration network between academics. Migrant and returning scientists tend to occupy more central and gatekeeper positions, and these positions play a significant role in determining the structure of collaborative networks among researchers in Uruguay. These results provide empirical evidence of the important contribution made by these actors, through the introduction of knowledge from outside the country and its dissemination at the local level.

References

Gould, R. V., & Fernandez, R. M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological methodology*, 89-126.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R*, (Vol. 65). New York, NY: Springer.

König, M. D., Battiston, S., Napoletano, M., & Schweitzer, F. (2011). Recombinant knowledge and the evolution of innovation networks, *Journal of Economic Behavior y Organization*, *79*(3), 145-164.

Robins, G., Pattison, P., Kalish, Y., y Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, *29*(2), 173-191.

Tomasello, M. V., Napoletano, M., Garas, A., & Schweitzer, F. (2017). The rise and fall of RyD networks, *Industrial and corporate change*, *26*(4): 617-646.

Can crowdsourcing rescue the social marketplace of ideas?

Taha Yasseri [1], Fil Menczer [2]

[1]School of Sociology, University College Dublin, Dublin, Ireland
[2]Observatory on Social Media, Indiana University, Bloomington, USA

***Keywords**: Homophily, Misinformation, Crowdsourcing, Community, Social Media*

How is it possible that some social web technology such as Wikipedia stand as the most successful example of collaborative and healthy information sharing, while the others, such as Twitter are blamed for epistemic chaos? To be sure, there are many important differences between Wikipedia and social media platforms, including design elements, business models, and user motivations and characteristics. However, a review of past research points to the network effects of content generation as a key to understanding how community-based moderation could rescue the social media marketplace of ideas, provided there is a serious intention by the commercial platforms to promote a healthier information environment.

The current social media approach of hiding conflict through self-selection, unfollowing, removal of content, and account bans neither prevents nor mitigates conflict. Some researchers suggested exposing users to counter-attitudinal content, a positive algorithmic bias designed to break the bubbles. However, experiments show that partisan users become more entrenched in their beliefs once they are exposed to opposing views [1]. What is effective, in contrast, is sharing personal experience and —if we have learned one thing from the Wikipedia experience— collaborative interaction. Such collaboration in the context of social media could be aimed at tackling misinformation and community policy violations. Recent experiments suggest that crowdsourced layperson judgments can be effective at identifying misinformation [2]. Such a community approach could scale up fact-checking and moderation practices while mitigating both misinformation and polarization.

Following this line of argument, Facebook announced a community review program in December 2019 and Twitter launched a community platform to address misinformation4 in January 2021. Here we focus on Twitter's platform, called Birdwatch, for which some preliminary data is available. In the current Birdwatch implementation, a member of the group of reviewers (selected by Twitter based on undisclosed criteria) can add a note to a tweet that they find "misinformed or potentially misleading." A note provides some information selected from predefined values about the tweet (misleading factual error, misleading satire) as well as some free text where the reviewer can comment and link to external sources. Then other reviewers express their agreement or disagreement with the existing notes through additional annotations such as helpfulness and informativeness. Ultimately, notes produced by reviewers will become visible next to the corresponding tweets based on the support/opposition they have received from other reviewers.

We analyzed Birdwatch helpfulness ratings as of February 2022 — 189,744 ratings of 17,888 notes by 7,884 reviewers. We observed evidence of a highly balanced network with two well-separated clusters where reviewers agree with those in the same group and disagree with those in the opposite group. In fact, of the pairs of reviewers with reciprocal ratings, 71% are consistent in that they both rate each other helpful (in the same cluster) or not helpful (in different clusters). Furthermore,

despite 35% negative ratings, we found that only 22% of triads of reviewers with reciprocal ratings are structurally imbalanced, i.e., inconsistent with all three reviewers agreeing with each other (in the same cluster) or with two reviewers in agreement with each other (in the same cluster) and in disagreement with the third (in the other cluster). Fig. 1 offers visual confirmation of these findings by mapping the network of Birdwatch reviewers. An edge between two nodes represents reciprocal ratings that indicate agreement on average. The polarization among Birdwatch reviewers mimics the one observed among generic Twitter users.

It is unlikely that this polarization is a reflection of objective arguments; rather, it merely represents the political affiliations of the reviewers. Analysis of the notes confirms that users systematically reject content from those with whom they disagree politically [3]. One might argue that the population of Birdwatch reviewers is less homogenous than that of Wikipedia editors. While this may be true, a polarized crowd can be even more effective in producing high-quality content compared with a homogenous team [4]. The missing ingredient, however, is collaboration: reviewers of opposing opinions currently do not have to reach a consensus. The design of Birdwatch will have to be modified to enforce collaboration rather than competitive behaviour; robustness to competition is as critical as resistance to coordinated manipulation. Wikipedia teaches us that community rules can enforce such norms.

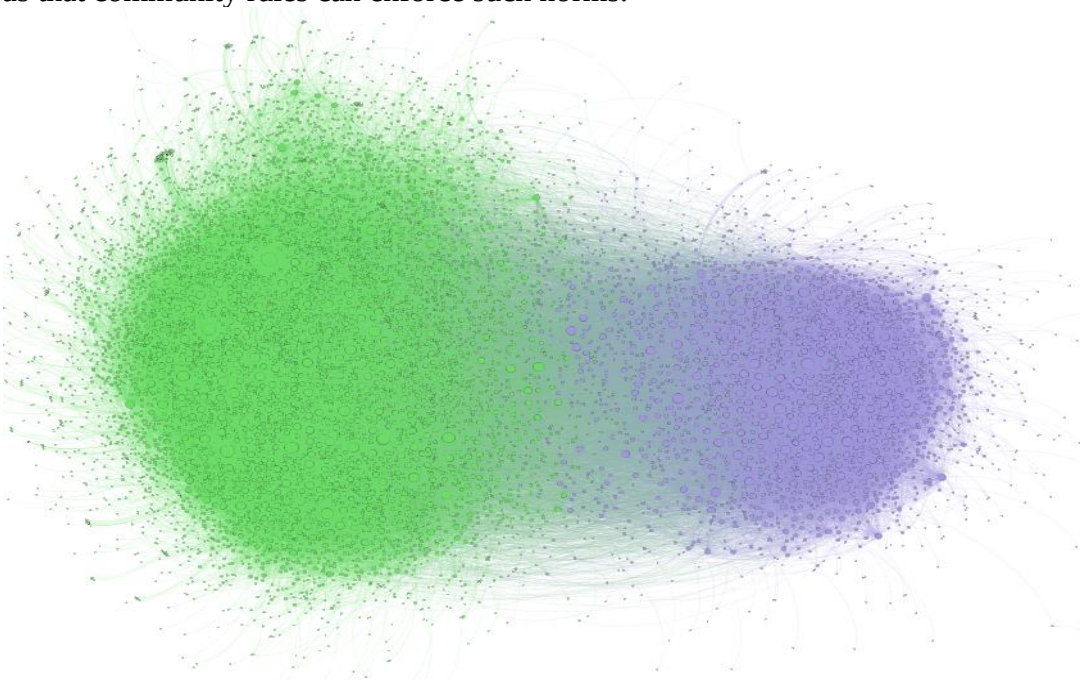


Fig. 1: The network structure of Birdwatch users and their positive ratings of each other's notes. Node colours are determined by a community detection algorithm and node size indicates the number of interactions.

[1] Kubin, E. et al. 2021. Personal experiences bridge moral and political divides better than facts. *PNAS*. 118, 6 (Jan. 2021), e2008389118.
[2] Nikolov, D. et al. 2021. Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *HKS Misinformation Review*. 1, 7 (Feb 2021).
[3] Allen, Jennifer et al. 2021. Birds of a Feather Don't Fact-check Each Other. Preprint PsyArXiv.
[4] Shi, F. et al. 2019. The wisdom of polarized crowds. Nature Human Behaviour. 3, 4 (Mar. 2019), 329–336.

Centrality-Based Ranking of Paired Comparison

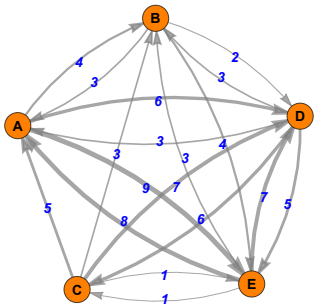
Yang Li *

Department of Mathematical Sciences, Florida Atlantic University
Boca Raton, FL 33431, USA

Keywords: Paired comparison, sports, ranking, ordering

Ranking players or teams is always an interesting topic in sports. For instance, the ATP (Association of Tennis Professionals) tour has more than one thousand active professional players participating in sponsored tournaments around the world. In one season, each player can play against at most dozens of other players, only a small subset of the whole cohort, thus creating a super-sparse ranking network system. It is crucial to extract the intertwined relationships and rank all players based on the limited information available.

Suppose that there are n players or teams who compete in multi-player games or sports. The number of games played between player i and player j is $n_{ij} \geq 0$. Out of these n_{ij} games, player i wins a_{ij} times and player j wins $a_{ji} = n_{ij} - a_{ij}$ times. The win-loss record can be represented by a directed weighted network G whose adjacency matrix $\mathbf{A} = \{a_{ij}\}$. In general $\mathbf{A}$ is asymmetric with zeroes on the diagonal. Figure 1 shows an example of G and $\mathbf{A}$ for a small data set with $n = 5$. The direction of the arrows runs from the winner to the loser with a weight being the frequency of wins. For example, player A defeats player B four times and loses three times.



(a)

$$\mathbf{A} = \begin{pmatrix} 0 & 4 & 0 & 6 & 9 \\ 3 & 0 & 0 & 2 & 4 \\ 5 & 3 & 0 & 7 & 1 \\ 3 & 3 & 6 & 0 & 5 \\ 8 & 3 & 1 & 7 & 0 \end{pmatrix}$$

(b)

Figure 1: Game record (a) and the corresponding adjacency matrix (b). The widths of the edges are proportional to the weights.

For an undirected unweighted network, the eigenvector centrality of a given node is proportional to the sum of the centralities of its neighbors such that $x_i = \lambda \sum_{j=1}^{n} C_{ij} x_j$ where $C_{ij} = 1$ if nodes i and j are connected, and zero otherwise [Bonacich, 1987]. The solution is a list of the centrality measures which can be used to describe the significance of the nodes in the network.

For the directed weighted network in the current situation, a generalization of the eigenvector centrality is defined

*Email: yangli@fau.edu

as

$$x_i = \lambda \sum_{j \in \mathrm{Ne}(i)} \frac{a_{ij}}{a_{ij} + a_{ji}} x_j, \tag{1}$$

where x_i is the *score* of player i and $\mathrm{Ne}(i)$ is the set of its neighbors. The basic idea in (1) is score swapping. Two players i and j will exchange some of their scores based on their win-loss record. Player i gains a proportion $\frac{a_{ij}}{a_{ij}+a_{ji}}$ of player j's score, and meanwhile transfers $(1 - \frac{a_{ij}}{a_{ij}+a_{ji}})$ of its own score in return. The factor $\frac{a_{ij}}{a_{ij}+a_{ji}}$ is the estimated probability that player i is preferred to player j in a single game [Bradley and Terry, 1952]. Note that the swapped scores are not symmetric. The score increase of a stronger player (with larger x_i) by defeating an underdog is less than its loss if it loses the game. Similarly, a player acquires more by defeating a stronger opponent than a weaker one.

There is a drawback of the definition in (1). Let us consider two players k and l. Player k loses all games it has played. In other words, $a_{kj} = 0$ for all $j = 1, \ldots, n$ and the kth row in the adjacency matrix $\mathbf{A}$ contains only zeroes. It is easy to see that its score x_k must be zero since all coefficients $\frac{a_{kj}}{a_{kj}+a_{jk}}$ are zero. On the other hand, player l wins all games against player k but loses all other games. Still, player l should have a zero score because either $a_{lj} = 0$ for $j \neq k$ or $x_k = 0$ and thus $x_l = 0$ by (1). This does not make much sense in practice [Newman, 2018]. In fact, being able to compete against other players is an affirmation for its skill and capability. Not many players have the chance of playing against top players in sports. If they do, it is very likely that they are participating in some top-notch tournaments which implies that their skills must be at a relatively high level. Moreover, every player should get some credit for being able to compete against other players. One should gain something even though a game is lost. As a remedy, we propose to generalize (1) as

$$x_i = \lambda \sum_{j \in \mathrm{Ne}(i)} \frac{a_{ij} + \alpha}{a_{ij} + a_{ji} + 2\alpha} x_j + \beta. \tag{2}$$

The newly added term β is the constant extra amount that every player receive, similar to the counterpart in the Katz centrality [Katz, 1953]. This term ensures that all players have a positive score, even though they lose all games they have played, and thus pass it along to others in the network. The new term α in the coefficient represents the gain from playing against other players.

Equation (2) can be written in the matrix form $\mathbf{x} = \lambda \widetilde{\mathbf{A}} \mathbf{x} + \beta \mathbf{1}$ where $\widetilde{\mathbf{A}}$ is an $n \times n$ matrix with elements $\widetilde{A}_{ij} = \frac{a_{ij}+\alpha}{a_{ij}+a_{ji}+2\alpha} = \frac{a_{ij}+\alpha}{n_{ij}+2\alpha}$ if $n_{ij} > 0$, and zero otherwise; and $\mathbf{1} = (1, 1, \ldots, 1)$ is an n-vector of ones. The solution is $\mathbf{x} = \beta(\mathbf{I} - \lambda \widetilde{\mathbf{A}})^{-1} \mathbf{1}$ conditioned on the fact the inverse exists. The overall coefficient β is not important since we typically only care about the relative magnitude (rank) of scores and a multiplication of an overall constant will not change the ranks. We can use $\beta = 1$ for simplicity. Additionally, in order for the inverse to exist, λ must be less than the reciprocal of the largest (most positive) eigenvalue of $\widetilde{\mathbf{A}}$ [Newman, 2018].

Some generalizations can be accommodated for more realistic situations. First of all, if a tie/draw is allowed as in chess and soccer games, a traditional way is to assign a half win to both players. If a weaker player draws a game against a stronger player, it is considered a "victory" for the underdog who then gains more from the score swapping than its opponent does. Additionally, different games may have distinct importance. For instance, winning a

Grand Slam final should bring more glory and honor than winning a qualifying match. a_{ij} needs to be replaced by w_{ij}, the total weight of winning games of player i against player j. Equation (2) becomes

$$x_i = \lambda \sum_{j \in \mathrm{Ne}(i)} \frac{w_{ij} + b_{ij}/2 + \alpha}{w_{ij} + a_{ji} + b_{ij} + 2\alpha} x_j + \beta.$$

where b_{ij} is the number of ties between player i and player j.

In the simulation study, we assume that there are 50 players in a multi-player game network. Their skill levels u_i $(i = 1, \ldots, 50)$ are equally spaced on the interval of $[1, 5]$. Higher values of skill level indicate stronger players. The strongest player is player 1 with $u_1 = 5$ and the weakest player is player 50 with $u_{50} = 1$. The number of games n_{ij} completed between a pair of players i and j follows a Poisson distribution with mean 8. The probability that player i wins any game against player j is $p_{ij} = u_i/(u_i + u_j)$, and the results of all games are independent. The simulation is carried out 200 times. In each repetition, all players are ranked and we tabulate all 200 ranked lists and summarize using the heat map in Figure 2. The darkness of the grid represents the frequency of the pairs of true and estimated ranks.

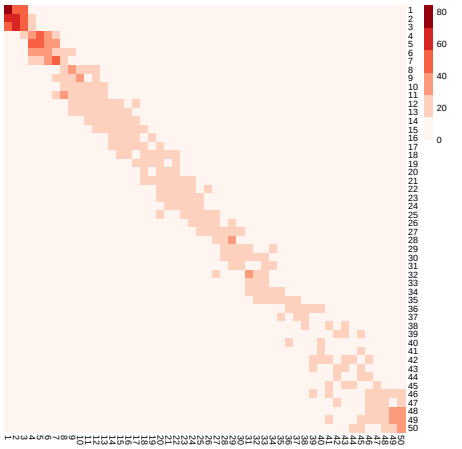


Figure 2: A heat map summary of 200 simulation repetitions. The horizontal axis is the true ranking of the players. The vertical axis is the estimated ranking.

A real data analysis is also performed on the records of ATP Tour in 2019 [Sackmann, 2022]. The results are compared to the merit-based method used by ATP. Other applications include the identification of disease-associated top (hub) genes in RNA sequencing-derived gene-gene interaction data (e.g., correlation or weights).

References

Phillip Bonacich. Power and centrality: a family of measures. *American Journal of Sociology*, 92:1170–1182, 1987.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

Mark Newman. *Networks*. Oxford University Press, 2 edition, 2018.

Jeff Sackmann. ATP tennis rankings, results, and stats. https://github.com/JeffSackmann/tennis_atp, 2022. Accessed: 2022-10-08.

Complex networks coarse-graining by Laplacian Renormalization Group

Tommaso Gili[1], Guido Caldarelli[2], Pablo Villegas[3], Andrea Gabrielli[3,4]
[1]Networks Unit, IMT - School for Advanced Studies, Lucca, Italy
[2]Dipartimento di Scienze Molecolari e Nanosistemi, Università "Ca' Foscari", Venice, Italy
[3]Enrico Fermi" Research Center – CREF, Rome, Italy
[4]Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Keywords: coarse-graining, renormalization group, laplacian matrix, density operator, diffusion processes

Complex networks usually exhibit a rich architecture organized over multiple intertwined scales. Information pathways are expected to pervade these scales reflecting structural insights that are not manifest from analyses of the network topology. Moreover, small-world effects correlate with the different network hierarchies complicating the identification of coexisting mesoscopic structures and functional cores. We present a communicability analysis of effective information pathways throughout complex networks based on information diffusion to shed further light on these issues. This will lead us to a formulation of a new and general renormalization group scheme for heterogeneous networks. The Renormalization Group (RG) is the cornerstone of the modern theory of universality and phase transitions, a powerful tool to scrutinize symmetries and organizational scales in dynamical systems. However, its network counterpart is particularly challenging due to correlations between intertwined scales. To date, the explorations are based on hidden geometries hypotheses. Here, we propose a Laplacian RG diffusion-based picture for complex networks, defining the supernodes concept à la Kadanoff, the equivalent momentum space procedure à la Wilson for graphs, and applying this RG scheme to real networks in a natural and parsimonious way (see Fig.1).

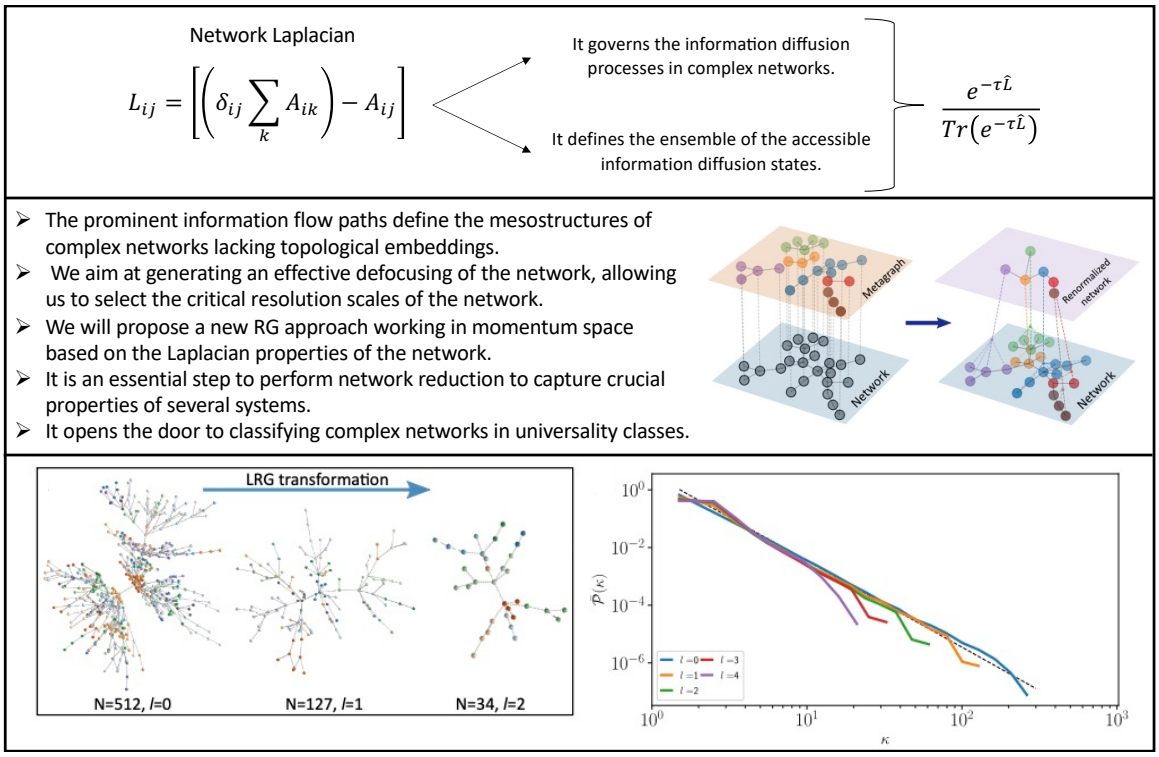


Fig.1. Coarse graining by Laplacian Renormalization Group: the case of Barabási-Albert networks.

References [1] *P. Villegas, A. Gabrielli, G. Caldarelli, T. Gili,* ***Laplacian paths in complex networks: Information core emerges from entropic transitions.*** *P. Villegas, A. Gabrielli, F. Santucci, G. Caldarelli, T. Gili, Phys. Rev. Research 4, 033196 (2022). [2] P. Villegas, T. Gili, G. Caldarelli, A. Gabrielli,* ***Laplacian Renormalization Group for heterogeneous networks****, https://doi.org/10.48550/arXiv.2203.07230 (2022).*

Complexity emerges in measures of the marking dynamics in football games

A. Chacoma[1,2], O.V. Billoni[1,2], M. N. Kuperman[3,4]

1 Instituto de Física Enrique Gaviola (IFEG-CONICET), Ciudad Universitaria, 5000 Córdoba, Argentina.
2 Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina.
3 Instituto Balseiro, Universidad Nacional de Cuyo, R8402AGP Bariloche, Argentina.
4 Centro Atómico Bariloche and CONICET, R8402AGP Bariloche, Argentina

This work was published as a regular article in Physical Review E (Featured in the physics)

https://journals.aps.org/pre/abstract/10.1103/PhysRevE.106.044308

In this work, we aim to study the marking dynamics using network science. To do so, we survey a database containing the coordinates of the players in the field at each second of three professional games. With this information, we define a bipartite graph where the nodes are the players of both teams, but the connections can be only between opponents.

To establish the connections in our networks, we will use the euclidean distance of the players in the field since the opponents' closeness is strictly related to the marking. This particular type of graph is known as proximity network and has been widely used to study multiple phenomena in complexity science.

Summarising, we observed that the proximity network evolves following the marking dynamics, exhibiting oscillating periods of high defragmentation and high clusterization.
To characterise this phenomenon, we calculated the heterogeneity parameter (κ) and found that the system evolves in a regime similar to a transition in percolation theory.
Since the system is far from the thermodynamic limit, we cannot frame our results in the theory of phase transitions. Our observations, however, evidence the emergence of complexity in the marking dynamics.

We were able to study this complex behaviour by analysing the temporal structure of the time series of κ. We found the presence of anti-persistency and self-similarity, which we characterised by uncovering a scaling law in the average shape of the fluctuations, see Fig. 1.

Lastly, we proposed a model to simulate the players' motion on the field. From simulations, we obtained the evolution of a synthetic proximity network that we analysed with the same methodology we used in our analysis of the empirical data. Remarkably, the model showed a good performance in recovering the statistics of the empirical trajectories; and, consequently, the statistics of the temporal structure of the parameter κ.

In conclusion, we can state that the correlations observed in the proximity network associated with the marking dynamics could be related to the high level of coordination required to keep running the tactical system. In this sense, our framework based on proximity networks allows us to observe that at each game challenge, the entire team will

proceed in coordination to give a response. They will tend to react optimally, according to the training precepts received. Therefore, it is expected that, in similar situations, they will produce equivalent responses. In our framework, these responses are encoded in the proximity networks as recurrent configurations and yield the memory effects we observe in the evolution of the heterogeneity parameter.

Moreover, the presence of correlations reveals the players are strongly connected. These connections drive the team to behave flexible and adaptable to stimuli, something crucial for the development of the game. We can compare this "state of alert" of the teams with what occurs with bird flocks or fish shoals, in which connections among the individual make the group stronger to avoid predators. The difference between these cases and the dynamics of a football team relies on the cognition capabilities required to achieve this level of organisation among the group's individuals.

The emergence of complexity in the game of football is somewhat similar to that observed in a living system. In these systems, when the delicate equilibrium between inhibition and promotion, cooperation and competition, is unbalanced, something abnormal occurs. This effect is observed, for example, in the appearance of cancer cells, in diseases of the nervous system, in diseased mitochondria, etc. When the complexity of the system is lacking, its functioning is severely damaged. Analogously, in the case of football dynamics, the lack of complexity would be related to low level played games. Therefore, our framework provides a tool that can help to detect a lack of performance in the teams.

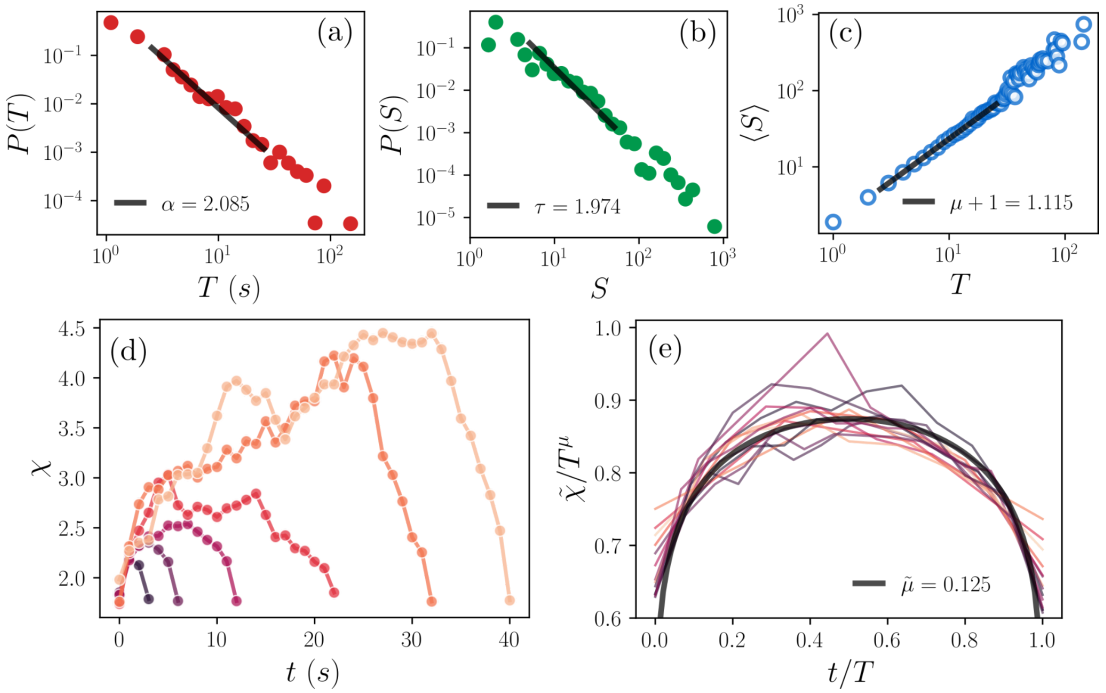


Fig. 1 Self-similarity in the series of κ. (a and b) Probability distributions of avalanche lifetime P(T) and avalanche size P(S). (c) Relation between avalanches lifetime T and the mean value of the size of the avalanches S. (d) Several examples of avalanches with a different lifetime. (e) Collapse of the avalanches produced by rescaling. Black solid lines in (a)–(c) and (e) show the result of nonlinear fit in the drawn regions.

Cooperation in costly-access environments

Hugo Pérez-Martínez[1], Carlos Gracia-Lázaro[2], Fabio Dercole[3], Yamir Moreno[2,4,5]

[1]Department of Condensed Matter Physics, University of Zaragoza. Spain.
[2]Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza. Spain.
[3]Department of Electronics, Information, and Bioengineering, Politecnico di Milano. Italy.
[4]Department of Theoretical Physics. University of Zaragoza. Spain.
[5]CENTAI Institute, Turin. Italy.

Understanding cooperative behavior in biological and social systems constitutes a scientific challenge, being the object of intense research over the past decades. Many mechanisms have been proposed to explain the presence and persistence of cooperation in those systems, showing that there is no unique explanation, as different scenarios have different possible driving forces. In this regard, the Evolutionary Game Theory provides a fruitful theoretical framework for studying cooperative behavior, including cooperation in structured populations. Within this framework, the Prisoner's Dilemma (PD) constitutes the most representative and widely studied game for modeling cooperative behavior evolution.

In this work, we propose a model to study situations where the willingness to participate in a cooperative setup involves an access cost (besides the cost associated with cooperation), even if the other potential participant player refuses the interaction. The motivation is to study those scenarios where the access to the interaction place (whether physical or virtual) entails an expense, such as transport costs, entry fees, or time investment, which can be avoided by refusing interaction. The proposed model corresponds to a reciprocal Donation Game with voluntary and costly participation. By imposing a participation fee, we break the symmetry of the Voluntary PD through a payoffs difference between the player that refuses to interact (the abstainer) and her counterpart (the attendant, i.e., cooperator or defector). The proposed two-person game, hereafter Costly-Access Prisoner's Dilemma (CAPD), has three strategies: cooperation, defection, and abstention. While abstention does not involve any payoff (neither benefit nor fee), defection and cooperation entail a participation fee besides the cooperation cost associated with the latter. Note that, in the proposed CAPD, the players willing to participate in the underlying PD pay the participation fee, regardless of whether PD takes place or not. Conversely, in the Voluntary PD, the participation fee is paid only if the PD takes place. We can interpret the CAPD as a risky version of the Voluntary PD, as showing up to participate involves a risk that breaks the symmetry of the Voluntary PD.

A mean-field approach shows that, in well-mixed populations, the dynamic always leads the system to abstention. However, depending on the return parameter, numerical simulations in structured populations display an alternating behavior between mono-strategic, multi-stable, and coexistence phases. This behavior is fully explained through a theoretical analysis of the strategic motifs, the transitions being determined by the change in stability of those motifs.

The Model

Let be a population of n agents –the players– endowed with a network structure. The interaction between any pair of agents takes place through a PD with voluntary and costly participation. Specifically, agents are allowed to adopt one of the three available strategies: cooperation (C), defection (D), and abstention (A). Each agent takes one of the above strategies when playing with all her neighbors. If the agent decides to abstain, she does not pay nor receive anything. Otherwise, she must pay a participation fee t. Furthermore, cooperation has an additional cost $c = 1$ (the contribution) to the participation fee, while defection does not entail additional cost. The counterpart of a cooperator receives rc, i.e., her partner's contribution c multiplied by the enhancement factor r. Once two players decide to participate (none is an abstainer), this formulation of the PD is equivalent to a reciprocal Donation Game: in that game, each player is allowed to donate c; if she does, her counterpart receives rc.

Let σ_i be the strategy of player i; $\sigma_i^T = (1,0,0), (0,1,0), (0,0,1)$ for an abstainer, cooperator and defector, respectively. The payoff obtained by player i facing j is given by $\sigma_i^T M \sigma_j$, where M is the payoffs matrix:

$$M = \begin{pmatrix} 0 & 0 & 0 \\ -t & r-t-1 & -t-1 \\ -t & r-t & -t \end{pmatrix}$$

A cooperator or a defector who faces an abstainer must pay the participation fee t because she has presented himself to play, but since the game is not played (the abstainer does not appear), cooperator will not pay the cooperation cost $c = 1$. Furthermore, the highest payoff corresponds to a defector facing (and therefore exploiting) a cooperator, while the lowest to a cooperator facing a defector.

Each agent plays with the same strategy with all her neighbors. The payoff Π_i of an agent, i, will be the sum of those obtained when playing with all her neighbors. Once all the agents have played, they decide whether to keep their strategy in the next round or to change it. Specifically, each agent i randomly selects a neighbor j and compares their payoffs. Agent i adopts j's strategy with a probability given by:

$$P_{ij} = \frac{1}{1 + \exp(\frac{\Pi_i - \Pi_j}{T})} \quad . \tag{1}$$

Summary of results and discussion

In the absence of a network structure, the only evolutionarily stable strategy is abstention, which coincides with the only Nash equilibrium of the system, as defection is in the PD. We have studied the system through a mean-field approach, which resulted in the absence of inner fixed points, revealing the lack of coexistence stationary states. Furthermore, mean-field trajectories are such that any initial condition ends in a full-abstention state.

Extensive numerical simulations on graphs have shown that the network structure has a determining influence on the system dynamics. We have identified a series of strategic motifs whose stability thresholds determine the system behavior. The motifs' analysis allowed us to explain the different phases of the system and locate the critical values of the game return parameter r that demarcate their boundaries. Also, it explains the jumps in the average strategies frequencies and provides the values of r at which they take place. In particular, the system exhibits i) a first mono-strategic phase dominated by abstention, followed by ii) a multi-stable mono-strategic phase in which dynamics leads to one of the three absorbing states, iii) a three-strategies coexistence phase, and iv) for heterogeneous networks, a phase dominated by cooperation with a residual presence of defectors.

To conclude, by breaking the symmetry of the Voluntary Prisoners' Dilemma, we have presented a model that displays a rich phenomenology, with alternating stability switches driven by a single parameter. We believe this behavior will deserve the attention of physicists and mathematicians to be extrapolated to other systems.

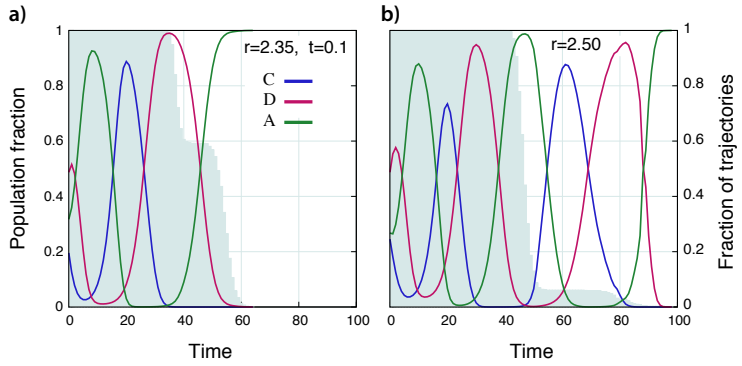


Figure 1: **Time evolution for an RRN.** Frequencies of the total fractions of each strategy as a function of time in a regular random network of $N = 10^4$ nodes and degree $k = 4$. Panel **a)** shows the results for $r = 2.35$, and **b)** for $r = 2.5$. Solid lines correspond to the average over 1000 simulations, and grey shadow to the fraction of trajectories remaining active at a certain time. $T = 0$ and $t = 0.1$.

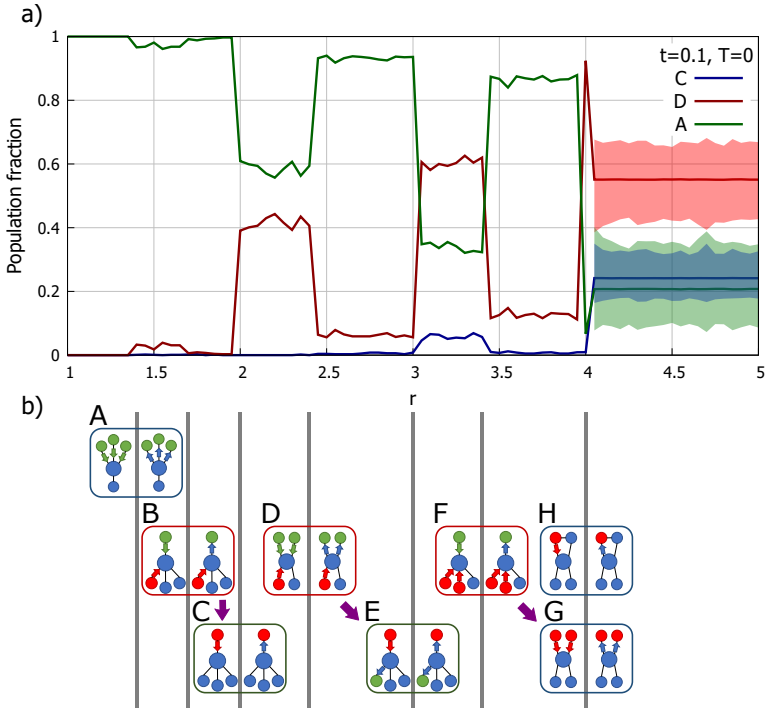


Figure 2: **Numerical results for an RRN, T=0. a)** Fraction of cooperators (C), defectors (D), and abstainers (A) versus the enhancement factor r in a RRN of 10^4 nodes, $k = 4$. **b)** Diagrams A to H correspond to the configuration transitions. Each diagram indicates a dominance transition for the corresponding motifs regarding the propagation or extinction of the central node. In these diagrams, neighbors of cooperators' D-neighbors are not cooperators. Grey arrows indicate the system's evolution when the central C-node invades an A-neighbor.

Bibliography. Hugo Pérez-Martínez, Carlos Gracia-Lázaro, Fabio Dercole, Yamir Moreno. Cooperation in costly-access environments. New Journal of Physics 24, 083005. 2022.

COVID-19 spreading under containment actions

LAFEC

Laboratorio de física estadística computacional

F.E. Cornes[a], G.A. Frank[b] , C.O. Dorso[a,c]

[a] Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón I, Ciudad Universitaria, 1428 Buenos Aires, Argentina

[b] Unidad de Investigación y Desarrollo de las Ingenierías, Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Av. Medrano 951, 1179 Buenos Aires, Argentina

[c] Instituto de Física de Buenos Aires, Pabellón I, Ciudad Universitaria, 1428 Buenos Aires, Argentina

We propose an epidemiological model that includes the mobility patterns of the individuals, in the spirit to those considered in Refs.[1-3]. We assume that people move around in a city of 120x120 blocks with 300 inhabitants in each block. The mobility pattern is associated to a complex network in which nodes represent blocks while the links represent the traveling path of the individuals (see Fig. 1). We implemented three confinement strategies in order to mitigate the disease spreading: 1) global confinement, 2) partial restriction to mobility, and 3) localized confinement. In the first case, it was observed that a global isolation policy prevents the massive outbreak of the disease. In the second case, a partial restriction to mobility could lead to a massive contagion if this was not complemented with sanitary measures such as the use of masks and social distancing. Finally, a local isolation policy was proposed, conditioned to the health status of each block. It was observed that this mitigation strategy was able to contain and even reduce the outbreak of the disease by intervening in specific regions of the city according to their level of contagion. It was also observed that this strategy is capable of controlling the epidemic in the case that a certain proportion of those infected are asymptomatic.

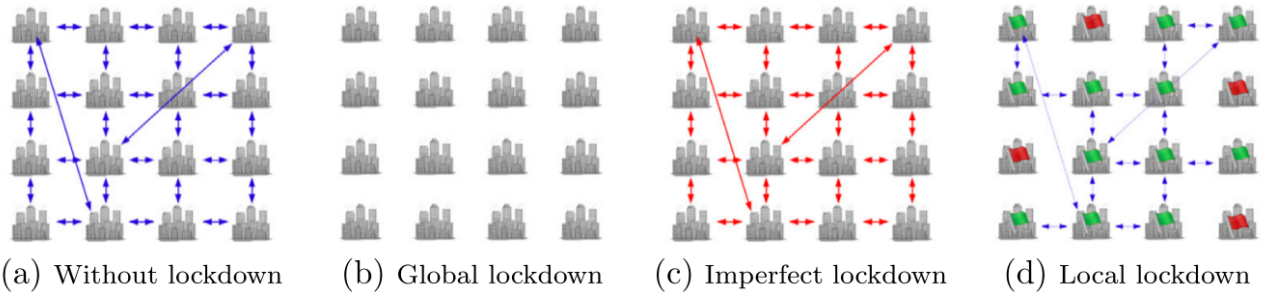


Fig. 1. (a–d) Schematic representation of different types of mitigation strategies (exemplified by a 4 × 4 grid). The lines represent the human mobility pattern between blocks. (a) People are allowed to evolve freely, equivalent to a "continuous activity" scenario. That is, there is no intervention during the epidemic. (b–d) Before a given day, individuals move freely from one block to another according to the mobility pattern. After that, (b) all blocks are

isolated, (c) the flow between blocks (represented by red arrows) is reduced but the mobility pattern is unaffected, (d) those blocks with a number of infected individuals greater than a certain threshold are isolated (labeled with a red flag). The other blocks remain linked to each other (labeled with a green flag). See text for more details.

References:

[1] A.D. Medus, C.O. Dorso, Diseases spreading through individual based models with realistic mobility patterns, 2011, arXiv:1104.4913.

[2] D.H. Barmak, C.O. Dorso, M. Otero, H.G. Solari, Dengue epidemics and human mobility, Phys. Rev. E 84 (2011) 011901.

[3] D. Barmak, C. Dorso, M. Otero, Modelling dengue epidemic spreading with human mobility, Physica A 447 (2016) 129–140.

[4] F.E. Cornes, G.A. Frank, C.O. Dorso, COVID-19 spreading under containment actions, Physica A 588 (2021) 12566

[5] F.E. Cornes, G.A. Frank, C.O. Dorso, Cycle strategy of lockdown and economic activity during the pandemic COVID-19 Anales Afa (2021) 150-156

Differences in nonlinear correlations between brain regions for patients with multiple sclerosis

M. Wątorek[1], M. Gawłowska[2], N. Golonka[2], J.K. Ochab[1,3], and P. Oświęcimka[1,4]

[1] Institute of Theoretical Physics, Jagiellonian University, Kraków, 30-348, Poland
marcin.watorek@uj.edu.pl
[2] Department of Cognitive Neuroscience and Neuroergonomics, Jagiellonian University, Kraków, 30-348, Poland
[3] M. Kac Complex Systems Research Center, Jagiellonian University, Kraków, 30-348, Poland
[4] Complex Systems Theory Department, Institute of Nuclear Physics Polish Academy of Sciences, ul. Radzikowskiego 152, 31–342 Kraków, Poland

Keywords: EEG, multiple sclerosis, nonlinear cross-correlations, multifractality

Multiple sclerosis (MS) is a chronic immune-mediated disease, the most common non-traumatic disorder of the central nervous system (CNS) [1]. The pathological hallmark of MS is demyelination and subsequent axonal degeneration that results in CNS lesions [2]. These neural alterations are present even in patients with early-stage MS [3]. Electroencephalography (EEG) can be used as a method to study secondary disease-induced changes in MS, such as cognitive impairment and other functional declines [4].

The multiscale methodology is one of the primary methods for studying complex systems and analyzing complex time series, which are undoubtedly EEG signals from the human brain [5].

In the investigation, a statistical analysis of the EEG data recorded during the resting state was performed for patients of the same age 30-40 with multiple sclerosis and the corresponding control group. The idea of the research was to examine differences in nonlinear cross-correlations, measured by the q-dependent detrended cross-correlation coefficient $\rho_q(s)$ [6], between brain regions represented by electrodes for various factors such as the duration of the disease, the stage of the disease, which is measured by the Expanded Disability Status Scale (EDSS), and medications administered during treatment. The results presented in Fig. 1 contain only the connection between two electrodes that exists, when the difference between two groups, for these electrodes, is statistically significant. The most significant differences are observed in case (b) - for group 1 patients who were being treated with Tecfidera®, and group 2 patients who were being treated with Interferon®. Differences between groups are also visible in case (c), where correlation matrices for patients in different stages of the disease (quantified by EDSS) from group 1 with EDSS > 1 and patients from group 1 with EDSS ≤ 1 are compared.

Furthermore, the fractal and multifractal properties of the EEG time series for 20 representative electrodes were also studied by using a multifractal detrended fluctuation analysis [7]. These measures can quantitatively describe the persistence and complexity of the considered time series [8, 9]. Looking at the differences in the width of the multifractal spectrum $\Delta\alpha$ - Fig. 2, it was possible to distinguish between the stage of the disease (quantified by EDSS -c) and the type of drug (b). There was almost no difference when the duration of the disease was taken into account. All observed differences were stronger in the phase of the experiment with closed eyes, which may be related to the delta waveform. It is also worth noting that the highest differences were observed in the time scale range $s = 200ms - 2500ms(5Hz - 0.4Hz)$, which corresponds to the delta wave.

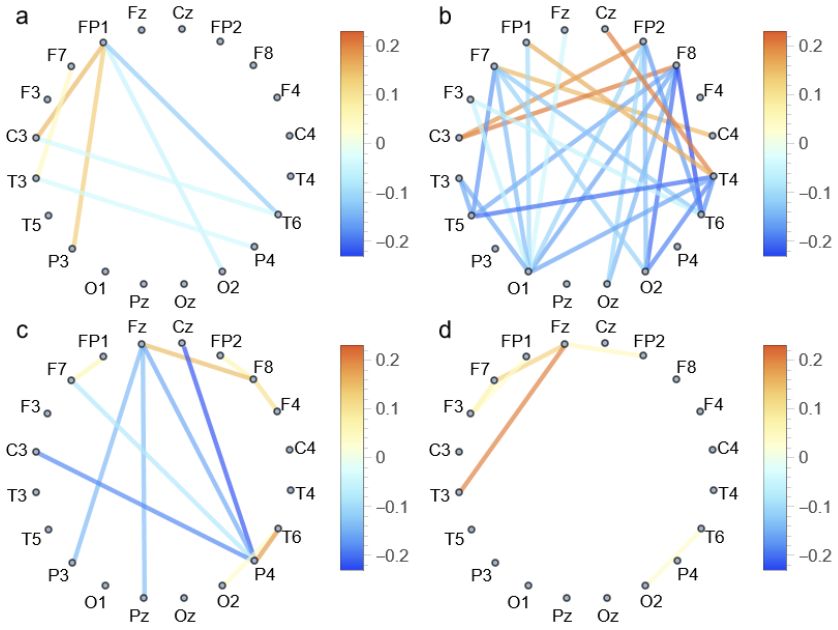


Fig. 1. Connections between 20 electrodes representing statistically significant differences between average correlation matrices ($\rho(q = 1, s = 200\,\text{ms}$) for various cases: (a) control and patients group, (b) patients 1 - Tecfidera® and patients 2 - Interferon® (c) patients 1 with EDSS > 1 and patients 1 with EDSS ≤ 1 (d) patients 1 with the time of disease longer than 7.5 years and patients 1 shorter than 7.5 years. In all cases, people between the ages of 30-40 and with closed eyes are considered.

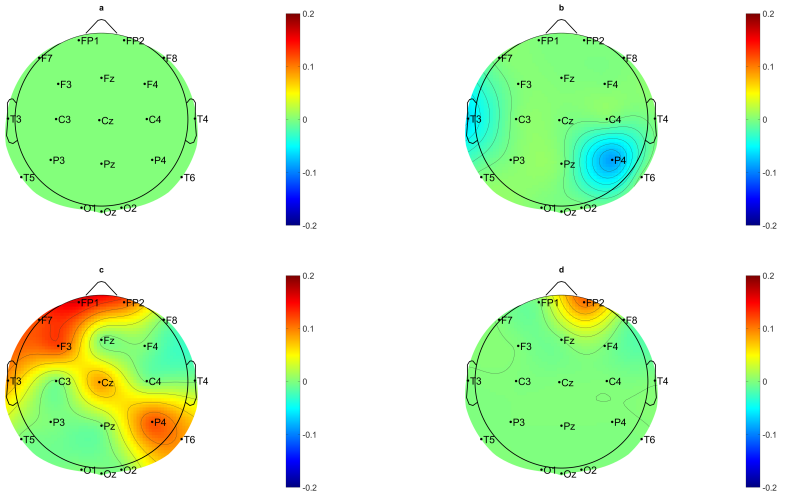


Fig. 2. Statistically significant differences between average multifractal spectra width $\Delta\alpha$ over 20 electrode areas for various cases: (a) control and patients group, (b) patients 1 - Tecfidera® and patients 2 - Interferon® (c) patients 1 with EDSS > 1 and patients 1 with EDSS ≤ 1 (d) patients 1 with time of disease longer than 7.5 years and patients 1 shorter than 7.5 years. In all cases people in the age 30-40 and with closed eyes are considered.

References

1. Dobson R, Giovannoni G. Multiple sclerosis–a review. European Journal of Neurology. 2019;26(1):27-40.
2. Bitsch A, Schuchardt J, Bunkowski S, Kuhlmann T, Brück W. Acute axonal injury in multiple sclerosis: correlation with demyelination and inflammation. Brain. 2000;123(6):1174-83.
3. Lucchinetti CF, Popescu BF, Bunyan RF, Moll NM, Roemer SF, Lassmann H, et al. Inflammatory cortical demyelination in early multiple sclerosis. New England Journal of Medicine. 2011;365(23):2188-97.
4. Vecchio F, Miraglia F, Porcaro C, Cottone C, Cancelli A, Rossini PM, et al. Electroencephalography-derived sensory and motor network topology in multiple sclerosis fatigue. Neurorehabilitation and Neural Repair. 2017;31(1):56-64.
5. Kwapień J, Drożdż S. Physical approach to complex systems. Physics Reports. 2012;515(3):115-226.
6. Kwapień J, Oświęcimka P, Drożdż S. Detrended fluctuation analysis made flexible to detect range of cross-correlated fluctuations.
7. Kantelhardt JW, Zschiegner SA, Koscielny-Bunde E, Havlin S, Bunde A, Stanley HE. Multifractal detrended fluctuation analysis of nonstationary time series. Physica A. 2002;316(1):87-114.
8. Drożdż S, Kwapień J, Oświęcimka P, Rak R. Quantitative features of multifractal subtleties in time-series. EPL (Europhysics Letters). 2010;88(6):60003.
9. Drożdż S, Oświęcimka P. Detecting and interpreting distortions in hierarchical organization of complex time series. Physical Review E. 2015;91:030902.

Digital cities and COVID-19: modeling the impact of non- pharmaceutical interventions

Jorge P. Rodríguez* [1,2,3] , Alberto Aleta [4] , Yamir Moreno [1,4]
[1] Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain
[2] Instituto Mediterráneo de Estudios Avanzados IMEDEA (CSIC-UIB), Esporles, Spain
[3]Centro Asociado Illes Balears, Universidad Nacional de Educación a Distancia (UNED), Palma, Spain
[4]ISI Foundation, Torino, Italy

*jrodriguez@imedea.uib-csic.es

Keywords

Non-pharmaceutical intervention, synthetic population, multilayer networks, agent-based models

Abstract

After a general lockdown that controlled the first wave (finishing on June 2020) of the COVID-19 outbreak in Spain [1], there were multiple non-pharmaceutical interventions at different geographical levels, from neighbourhoods to autonomous regions. This diversity of interventions and the different implementation dates hindered the characterization of their impacts in the outcome of the second wave (from June to December 2020), both in number of infections and number of deaths.
In order to quantify these impacts, we performed an exhaustive data gathering from multiple sources, extracting relevant features that allowed inferring social interactions described by six connectivity layers, in the context of the multilayer network formalism [2]. Specifically, we considered different layers representing the interactions in households, schools, working places, universities, nursing homes and community (Fig. 1). These multilayer networks, together with the available metadata (age and sex), enabled the creation of five synthetic cities (Barcelona, Valencia, Sevilla, Zaragoza, Murcia), where we modeled the first and second waves of the spread of COVID-19 inside the cities.
After calibrating our models with empirical data, we simulated new scenarios where we modified or omitted the interventions that were taken in the real scenario. The comparison between these counterfactuals and reality allowed us to quantify the impact of the most relevant non-pharmaceutical interventions, as well as made it possible to contrast responses, real and possible, in different locations.

Figure

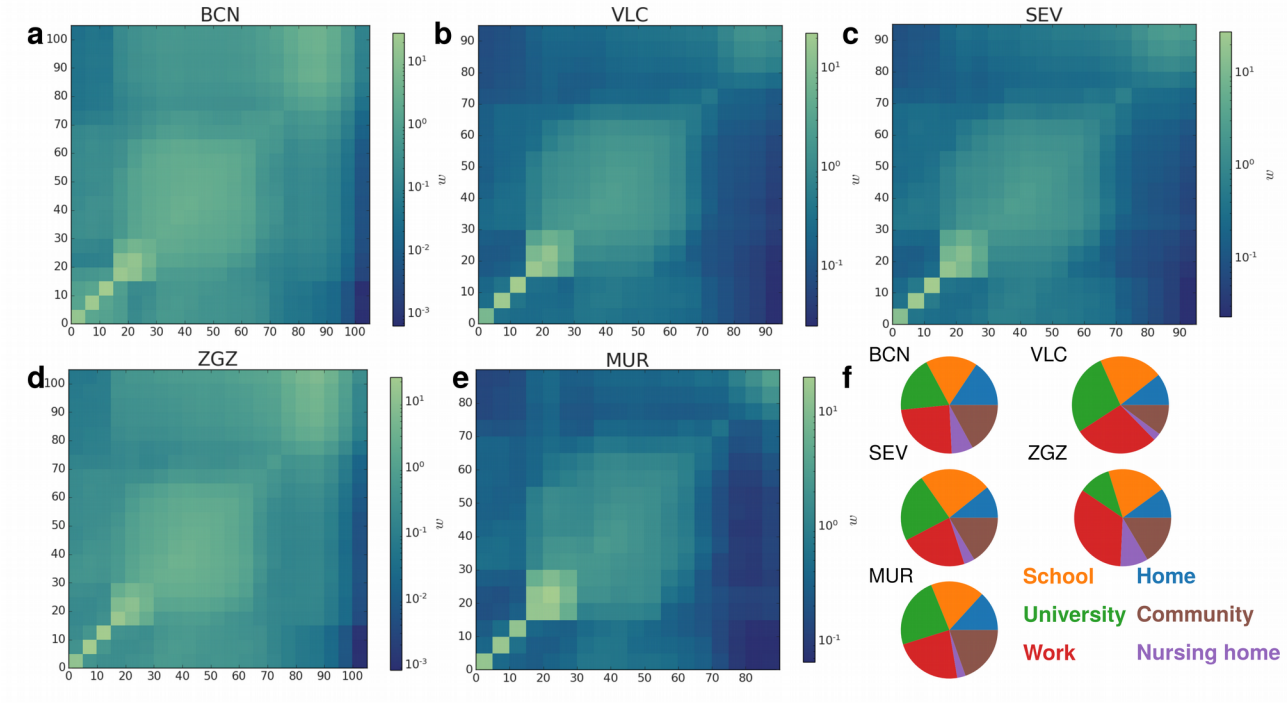


Figure 1. **a-e,** Inferred contact matrices by age for **a,** Barcelona (BCN), **b,** Valencia (VLC), **c,** Sevilla (SEV), **d,** Zaragoza (ZGZ), and **e,** Murcia (MUR). Rows and columns represent the pouplation of specific ages, and weights w represent the expected number of contacts for an individual from a specific row with individuals from each column. **f,** Fraction of links in each of the connectivity layers.

References

[1] Eguíluz, V. M., Fernández-Gracia, J., Rodríguez, J. P., Pericàs, J. M., & Melián, C. (2020). Risk of secondary infection waves of COVID-19 in an insular region: the case of the Balearic Islands, Spain. Frontiers in Medicine, 7, 563455.

[2] Aleta, A., & Moreno, Y. (2019). Multilayer networks in a nutshell. Annual Review of Condensed Matter Physics, 10, 45-62.

Discovering Semantic Relationships Using a Temporal Multiplex Network For a Context-Aware Filipino Wordnet

ROBI JEANNE A. BANOGON, De La Salle University
CHRISTINE G. DETICIO, De La Salle University
SHARMAINE S. GAW, De La Salle University
DANIELLE KIRSTEN T. SISON, De La Salle University
UNISSE C. CHUA, De La Salle University
CHARIBETH CHENG, De La Salle University
BRIANE PAUL V. SAMSON, De La Salle University

Additional Key Words and Phrases: Wordnet, Language networks, Multiplex networks, Community detection

ACM Reference Format:
Robi Jeanne A. Banogon, Christine G. Deticio, Sharmaine S. Gaw, Danielle Kirsten T. Sison, Unisse C. Chua, Charibeth Cheng, and Briane Paul V. Samson. 2022. Discovering Semantic Relationships Using a Temporal Multiplex Network For a Context-Aware Filipino Wordnet. *ACM Trans. Graph.* 37, 4, Article 111 (August 2022), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The low-resource and highly morphological setting of Philippine languages is a challenge in developing a word representation or a language model. Popular language resources such as the FilWordNet contains 10,344 synsets [Borra et al. 2010], which is considered small compared to other wordnets with synset sizes of over 50,000 [Khodak et al. 2017; Lam et al. 2019; Sand et al. 2017]. The current linguistic resources lack in rich semantic data that is crucial in most NLP tasks, and the fast-paced evolution and adaptation of Philippine languages make things even more difficult in creating a well defined language resources. As language evolves over time, new words and senses emerge in the everyday use of digital media. With the vast amount of data in many digital platforms that can represent different domains and varieties of words through time including changes in its semantic and syntactic forms, we aim to create word representations of the Filipino language that is temporal and context aware and store them in the expanded Filipino WordNet. As a language continues to evolve and show complexity, it may be represented as a language network to show the lexical, semantic, and syntactic features of a word [Seoane and Sole 2018].

We developed a data processing pipeline that maintains context-aware entries of Filipino and Philippine English words, their senses, and their semantic relationships, represented as semantic sets or semsets. Data from various sources

Authors' addresses: Robi Jeanne A. Banogon, robi_jeanne_banogon@dlsu.edu.ph, De La Salle University; Christine G. Deticio, christine_deticio@dlsu.edu.ph, De La Salle University; Sharmaine S. Gaw, sharmaine_gaw@dlsu.edu.ph, De La Salle University; Danielle Kirsten T. Sison, danielle_sison@dlsu.edu.ph, De La Salle University; Unisse C. Chua, unisse.chua@dlsu.edu.ph, De La Salle University; Charibeth Cheng, charibeth.cheng@dlsu.edu.ph, De La Salle University; Briane Paul V. Samson, briane.samson@dlsu.edu.ph, De La Salle University.

Manuscript submitted to ACM

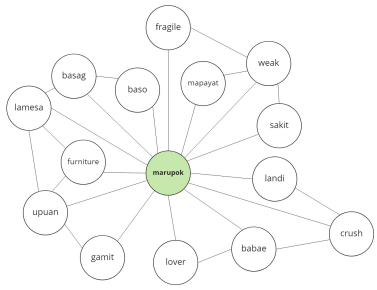


Fig. 1. An example of an ego network with nodes consisting of the ego and nodes that are at most one hop away from the ego.

(e.g. books, news, social media) were collected to build a corpus composed of 997,775,625 non-unique tokens. It is followed by building a temporal multiplex network, where layers correspond to sources, nodes represent unique words, and edges represent co-occurrence in a sentence or in one text [Sole et al. 2010]. Since a language can be used differently depending on the context it is used in, a co-occurrence language network can be reconfigured as a multiplex network with each layer corresponding to a different context [Kinsley et al. 2020]. The idea of a multiplex network can be further extended into a temporal multiplex network to model the evolution of a language across different time intervals in different contexts [Starnini et al. 2016]. Moreover, besides language being a complex system itself, using a network allows for the application of community detection algorithms and determining groups of word senses to form sense representations called semantic sets, shortened as semsets, with each set representing one or more semantic relationship among the senses.

Before inducing word senses through community detection, the multilayer network was flattened into an aggregated network composed of 1,074,594 nodes and 39,724,813 edges. Instead of inducing senses from the whole co-occurrence network, senses were induced individually on ego networks. An ego network was constructed for each target node, including only nodes within a maximum distance of one degree from the target node. An example of an ego network can be seen in Figure 1.

In order to filter out irrelevant nodes and edges, only alters that have edges with the ego that have weights above the top 90th percentile were retained, thus preserving nodes that frequently appear with the target word along with their edges. Afterwards, the ego of each ego network was dropped, thus retaining only the alters for community detection.

Three community detection algorithms were explored to create sense communities within each ego network: the Leiden algorithm, the Louvain algorithm, and the CW algorithm. The first two algorithms were performed using the leidenalg[1] library while the latter using the CDlib[2] library. Two quality functions were used for the Leiden algorithm, which are Constant Potts Model (CPM) with a resolution parameter, and Modularity. Communities that contained less than 10% of the total number of alters in the network were removed, as demonstrated in Figure 2. Each resulting community then represents a word sense for the ego.

With 4,557 seed words, word sense induction was finally performed using the Leiden community detection algorithm with modularity as its quality function and generated 11,548 word senses. New senses that only appeared recently, especially in social media, were also discovered, such as “awit”. A gold standard for 30 test words was created in order to validate the produced communities for community detection. The steps taken closely resembles the methodology

[1]https://leidenalg.readthedocs.io/en/stable/intro.html
[2]https://cdlib.readthedocs.io/en/latest/

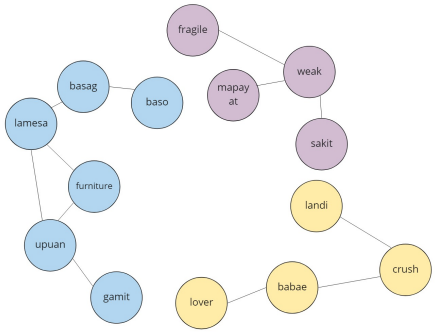


Fig. 2. Communities after clustering, resulting to 3 communities.

detailed by Bekavac & Ŝnajder [2016]. Annotators are tasked to cluster sentences, which are then compared to each other by constructing a probability matrix that shows the probability of two sentences appearing together in a community. The matrix is then used to construct an edgelist, then a network where community detection is performed. The resulting communities are the gold standard senses.

After which, 9,549 semsets were derived using the Jaccard index. Manual inspection of the induced semsets reveal synonyms, antonyms, hyponyms, hypernyms, and other semantic relationships, which can be used to update existing wordnets that only tackle synonymy. The pipeline was able to capture the dynamic and diverse usage of words across different mediums, giving researchers access to updated and relevant language resources. They will be made available to the general public and researchers through a public API.

REFERENCES

Marko Bekavac and Jan Šnajder. 2016. Graph-Based Induction of Word Senses in Croatian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 3014–3018. https://aclanthology.org/L16-1481

Allan Borra, Adam Pease, Rachel Edita, Roxas, and Shirley Dita. 2010. Introducing Filipino WordNet. In *Principles, Construction and Application of Multilingual Wordners: Proceedings of the 5th Global WordNet Conference*.

Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated WordNet Construction Using Word Embeddings. 12–23. https://doi.org/10.18653/v1/W17-1902

Amy C. Kinsley, Gianluigi Rossi, Matthew J. Silk, and Kimberly VanderWaal. 2020. Multilayer and Multiplex Networks: An Introduction to Their Use in Veterinary Epidemiology. *Frontiers in Veterinary Science* 7 (September 2020), 596. https://doi.org/10.3389/fvets.2020.00596

Khang Lam, Tuan To, Thong Tran, and Jugal Kalita. 2019. Improving Vietnamese WordNet using word embedding. *NLPIR 2019: Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, 110–114. https://doi.org/10.1145/3342827.3342854

Heidi Sand, Erik Velldal, and Lilja Øvrelid. 2017. Wordnet extension via word embeddings: Experiments on the Norwegian Wordnet. , 298–302 pages. https://www.aclweb.org/anthology/W17-0242

Luís Seoane and Ricard Sole. 2018. The morphospace of language networks. *Scientific Reports* 8 (July 2018). https://doi.org/10.1038/s41598-018-28820-0

Ricard Sole, Bernat Corominas-Murtra, Sergi Valverde, and Luc Steels. 2010. Language Networks: Their Structure, Function, and Evolution. *Complexity* 15 (July 2010), 20–26. https://doi.org/10.1002/cplx.20305

Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. 2016. Temporal correlations in social multiplex networks. *Scientific Reports* 7 (June 2016). https://doi.org/10.1038/s41598-017-07591-0

Distributed Self-healing under restricted communication: the effect on spatial damages

1st JAEHO KIM
Division of Transdisciplinary Sciences
Japan Advanced Institute of Science and Technology
Nomi, Japan
s2160002@jaist.ac.jp

2nd YUKIO HAYASHI
Division of Transdisciplinary Sciences
Japan Advanced Institute of Science and Technology
Nomi, Japan
yhayashi@jaist.ac.jp

***Keywords*—Self-healing, Robustness of connectivity, Spatially damaged network**

Many infrastructures of communication, transportation, or power-grid systems are represented by a common topological structure called scale-free (SF) [1]. Unfortunately, it is well-known that SF structure is extremely vulnerable against malicious attacks to high degree nodes. Moreover, these weak infrastructures are frequently damaged by natural and man-made disasters. However, the whole connectivity should be perpetually maintained to provide essential services for our daily life.

Therefore, a resilience-based system design has been attracted as an approach to overcome these problems. The concept of resilience includes not only recovering the original structure from disturbances but also reconstructing systems with adaptive capacity [2]. We emphasize that a reconstruction of a damaged network is more important for improving robustness of connectivity. Because, even when infrastructure systems are recovered to original structure, they still have the extremely weak SF structure. Thus, in order to reconstruct a damaged network into more robust one, we have proposed a distributed self-healing method especially based on enhancing loops [3].

The presence of loop structures in a network is a vital factor for the robustness. This is supported by that the network dismantling and decycling problems are asymptotically equivalent in random networks [6]. Here, the dismantling problem is finding the minimum set of nodes which removal makes a network fragmented into at most a given size, while the decycling problem is finding the minimum set of nodes which removal makes all loops from a network. When all loops are removed from a network, the network becomes a tree which is easily fragmented by a removal of few nodes. In other words, for constructing a robust network, it is important to make it hard to become a tree.

Therefore, we have adopted enhancing loops for generating robust structure. In fact, enhancing loops [7], [8] is more effective to generate an onion-like network than other methods such as increasing degree-degree correlations [4], [5]. Here, the onion-like network with positive degree-degree correlations has the optimal robustness against malicious attacks under a given degree distribution [4], [5], in contrast to SF structure.

In the proposed method [3], damaged nodes in a network communicate with other neighboring damaged nodes by exchanging short messages. To maintain a larger connectivity by healing, we assumed that each node of a network has a data set called local map storing identifiers of nodes within three hops from itself. Initially, the local map is defined as a set of candidates nodes for making new connections. Through communicating by some messages, such candidates are extended gradually. After that, damaged nodes which have the same candidates are connected by a ring. Then, the generated ring is enhanced by adding links between low degree nodes on the ring. For the rapid reconstruction, healing links are locally allocated to each node. Thus, it is assumed that nodes reuse some links emanated from attacked nodes. Since the number of reusable links corresponds to a damaged situation, a reusable rate of links is defined by a control parameter r_h which range is $0 < r_h \leq 1$ in our numerical simulation.

The effectiveness of our self-healing has been shown [3] through some investigations for typical infrastructure networks. The reconstructed networks by our method have high robustness against an intentional attack to important nodes with the highest degrees.

However, there is still considerable uncertainty with regard to the effect against other destructive attacks. Therefore, we consider a realistic scenario of disruption caused by spatial disasters. Since a destruction of a real network occurs as removing spatially grouped nodes by earthquakes or floods, we consider Localized Attack (LA) [9] as a malicious attack for reflecting such destruction into infrastructure networks.

In this paper, we have evaluated a network robustness against the attacks: High Degree Adaptive attack (HDA), and LA. HDA is removing highest degree nodes with recalculation of degrees for the networks. A part of connected component as neighbors of a randomly selected node is removed from the networks by LA.

For numerical evaluation of network robustness, we apply the robustness index [4] $R(q) = \frac{1}{N}\sum_{Q=1}^{N} S(Q)(\frac{1}{N} < R(q) \leq 0.5)$ for the reconstructed infrastructure networks after removing qN nodes by HDA, or LA. Here, N is the network size, and $S(Q)$ denotes a ratio of the largest connected component size after removing the number of Q nodes by HDA, or

LA. The index is investigated for the a typical infrastructure network: OpenFlight [10] and AS Oregon [11]. The values of $R(q)$ are averaged over 100 realizations for reconstructed networks.

We show robustness index against two attacks for the networks reconstructed after HDA in Fig.1(a) and (b), or after LA in Fig.1(c) and (d). In particular, we mainly present that the networks reconstructed after one attack also have high robustness against the other attacks.

As shown in each of Fig.1(a) and (c), our method has high values of R_{HDA} after HDA, and R_{LA} after LA. In other words, the reconstructed networks obtain a high robustness against the attack by which they have once been damaged.

Moreover, the reconstructed networks also have high values of robustness against the attack by which they have not been damaged. Fig.1(b) shows the values of $R_{\mathrm{LA}}(q)$ for the networks reconstructed after HDA, and Fig.1(d) shows the values of $R_{\mathrm{HDA}}(q)$ for the networks reconstructed after LA. To be more specific, the networks have the high values of $R_{\mathrm{LA}}(q)$ even for $r_h \leq 0.2$ (red, green, and blue lines in Fig.1(b)). The values of $R_{\mathrm{HDA}}(q)$ are also high especially for $r_h \geq 0.5$ (yellow, and purple lines in Fig.1(d)). However, when $q \geq 0.8$ for highly damaged situations by LA, the values of $R_{\mathrm{HDA}}(q)$ decreases rapidly. The reason of decreasing is considered as follows. Remember that each node can initially communicate with nodes in only three hops from it. Thus, when a huge hole is created by LA in a network, nodes cannot transfer messages to distant nodes. This problem leads to decreasing connectivity in our method. In other words, there is a trade-off between restriction of communication range and maintaining of connectivity. If infrastructures are spatially destroyed over 80%, our method may not work well. However, such highly damaged situations occur very rarely in real life. Even in that situations, it seems to be better not to heal networks but to construct a novel infrastructure. Therefore, we conclude that such trade-off does not become a limitation of our method and that our method is effective against localized attacks.

ACKNOWLEDGMENT

This research is supported in part by JSPS KAKENHI Grant Number JP.21H03425.

REFERENCES

[1] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, Classes of small-world networks, Proceedings of the national academy of sciences, 97(21), 11149-11152, 2000.

[2] C. Folke, Resilience: The emergence of a perspective for social–ecological systems analyses, Global environmental change, 16(3), 253-267, 2006.

[3] J. Kim, and Y. Hayashi, Distributed Self-Healing for Resilient Network Design in Local Resource Allocation Control, Frontiers in Physics, 272, 2022.

[4] C. M. Schneider, A. A. Moreira, J. S. Andrade Jr, S. Havlin, and H. J. Herrmann, Mitigation of malicious attacks on networks, Proceedings of the National Academy of Sciences, 108(10), 3838-3841, 2011.

[5] T. Tanizawa, S. Havlin, and H. E. Stanley, Robustness of onionlike correlated networks against targeted attacks, Physical Review E, 85(4), 046109, 2012.

[6] A. Braunstein, L. Dall'Asta, G. Semerjian, and L. Zdeborová, Network dismantling, Proceedings of the National Academy of Sciences, 113(44), 12368-12373, 2016.

[7] Y. Hayashi, and N. Uchiyama, Onion-like networks are both robust and resilient," Scientific reports, 8(1), 1-13, 2018.

[8] M. Chujyo, and Y. Hayashi, A loop enhancement strategy for network robustness," Applied Network Science, 6(1), 1-13, 2021.

[9] S. Shao, X. Huang, H.E. Stanley, and S. Havlin, Percolation of localized attack on complex networks, New Journal of Physics, 17(2), 023049, 2015.

[10] RA. Rossi, and NK. Ahmed, The network data repository with interactive graph analytics and visualization. AAAI, 2015.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 177–187, 2005.

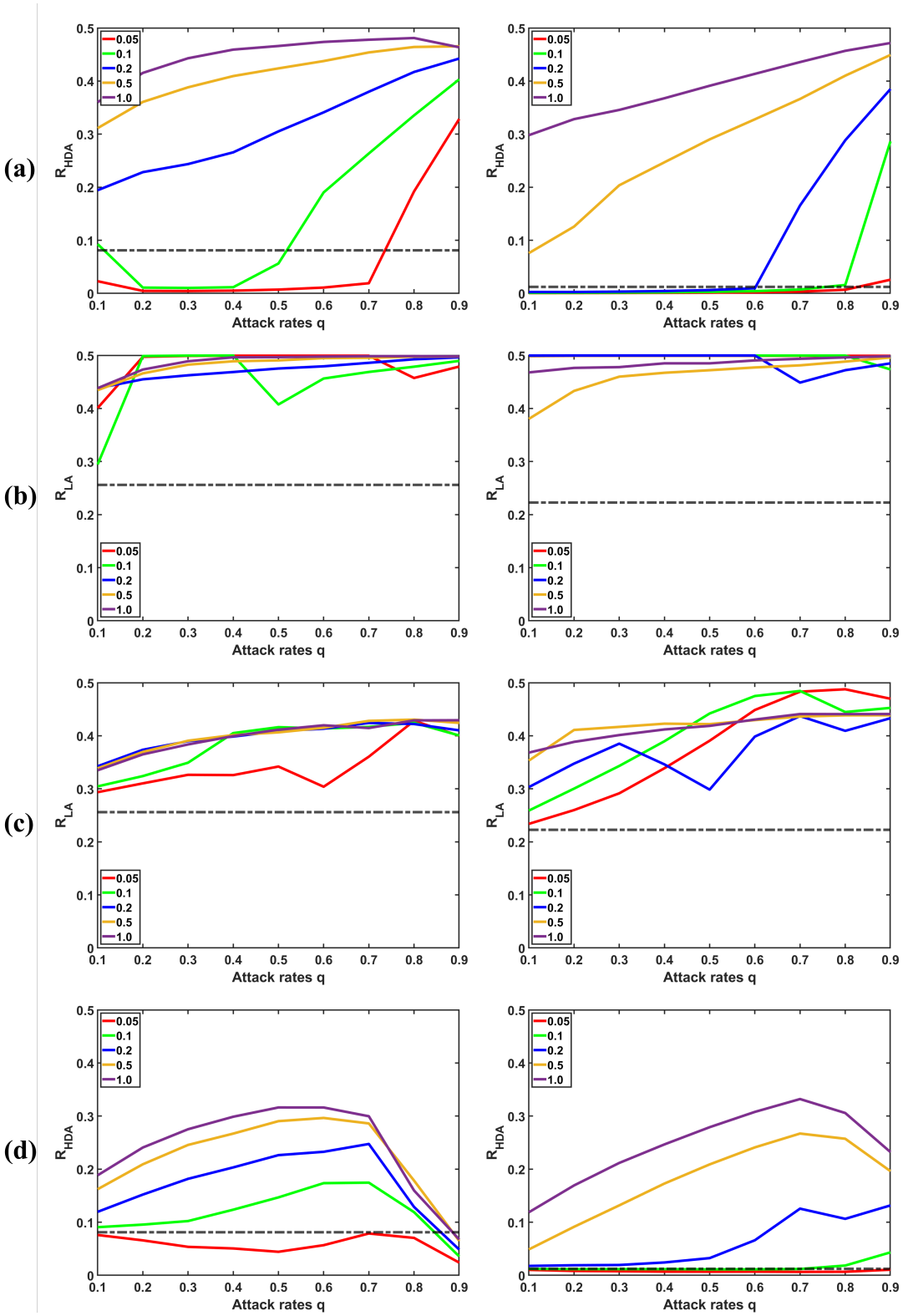


Fig. 1. Robustness of the reconstructed OpenFlight (left column) and AS Oregon (right column). (a) $R_{\mathrm{HDA}}(q)$ for the networks reconstructed after removing qN nodes by HDA. (b) $R_{\mathrm{LA}}(q)$ for them. (c) $R_{\mathrm{LA}}(q)$ for the networks reconstructed after removing qN nodes by LA. (d) $R_{\mathrm{HDA}}(q)$ for them. Colors correspond to several ratio r_h of reused links for healing. Black dot-dash line represents the robustness in the original network as a base-line.

Dynamics matter: A simulation framework to study diffusion processes on a Dynamic Product Space

Tobias Carreira-Munich
Instituto de Investigación en Ciencias de la Computación (ICC) - UBA/CONICET and Facultad de Ciencias Exactas y Naturales - UBA
Buenos Aires, Argentina
tcarreira@dc.uba.ar

Ezequiel Pecker-Marcosig
Instituto de Investigación en Ciencias de la Computación (ICC) - UBA/CONICET and Facultad de Ingeniería - UBA
Buenos Aires, Argentina
emarcosig@dc.uba.ar

Rodrigo Castro
Instituto de Investigación en Ciencias de la Computación (ICC) - UBA/CONICET and Facultad de Ciencias Exactas y Naturales - UBA
Buenos Aires, Argentina
rcastro@dc.uba.ar

ABSTRACT

We developed a simulation framework to experiment flexibly and robustly with the concept of Product Space (PS) including its extension as a Dynamic Product Space (DPS). We show that considering the DPS as a macroscopic emergent network property (at the global economy) impacts noticeably the diffusion processes occurring in the microscopic level (at individual countries), facilitating the production of new products in comparison to the classic (static) PS.

1 INTRODUCTION

Understanding the patterns of economic growth stands as a key tool for development planning and economic policy making. The discipline of Economic Complexity offers a framework to approach the study of interconnected economies as a Complex Adaptive System [8] that evolves in time driven by strongly non-linear dynamics.

A well known example is the *Product Space* (PS), an analytical framework proposed by Hidalgo et al. [7, 6, 5]. In the PS, a bipartite network is built based on worldwide foreign trade data, where countries are associated with the products they exchange in the global economy. The resulting network structure is then used as a proxy to measure the complexity of a given economy.

Namely, the distances between products in the PS can be used as indicators to assess the development potential of countries. This comes from the empirical observation that the PS networked structure impacts considerably on the development opportunities available to a given country. According to this framework, countries develop a so called *Revealed Comparative Advantage* (RCA)[1] in goods that are *closer* to those they already produce and export.

This framework opens up the possibility of conducting *what-if* prospective analyses, which are essentially simulations of a diffusion process on a dynamic network that can be used to assess potential paths of development for a country's productive structure. Such simulation studies on the PS have been attempted before [7].

However, even though these studies proved highly relevant, we argue that the reliability, replicability and transparency of the underlying simulation algorithms still remain under-attended requirements. Limitations that we have identified include:

(1) there is an unmet need to experiment flexibly and robustly with scenarios using of both, fixed and dynamic forms of the PS.
(2) there are no available open access tools to experiment with, and replicate, the simulation exercises.

In this work, we present results obtained by applying a new formal simulation framework to study diffusion processes over the PS to overcome these downsides.

2 METHODOLOGY

We begin by following the methodology explained in [7] to simulate diffusion in the PS, in an effort to replicate the results obtained there. Our numerical results appear, by the naked eye, quite similar to those reported in [7]. Yet, a precise numerical comparison is not possible since neither the original simulation algorithm nor its results are publicly available (our requests for data access have not yet been satisfied).

To build the PS we used public datasets of the world trade flows ranging from 1998 to 2000, classified using the Standard International Trade Classification rev. 4 (SITC-4) with 775 products and 190 countries, generated by the National Bureau of Economic Research (NBER) project (led by R. Feenstra [2])[1]. This way, we can easily reproduce simulations for any listed country and therefore perform comparisons between countries.

We define the *proximity matrix* Φ with elements $\Phi_{i,j}$ representing the *proximity between products i and j* in the PS. Each entry is defined by the minimum conditional probability of some country having RCA>1 in product i given that the same country already has an RCA>1 in product j:

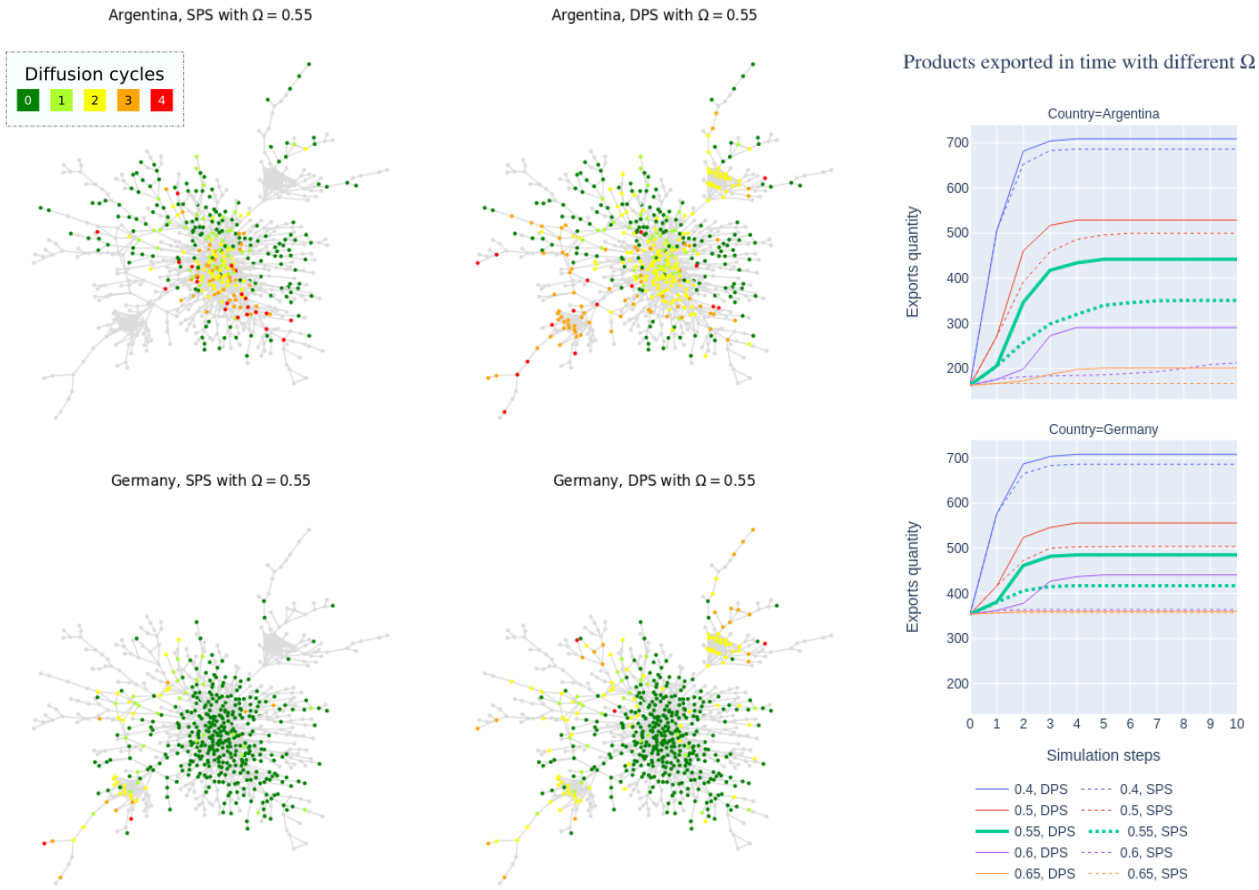
$$\Phi_{i,j} = \min\{P_{i,j}, P_{j,i}\}, \text{ with } P_{i,j} = \frac{\sum_c M_{c,i} M_{c,j}}{\sum_c M_{c,i}} \tag{1}$$

where $P_{i,j}$ is the conditional probability (or frequency) of producing a product i provided that product j is already being produced (and vice versa for $P_{j,i}$). As we mentioned before, this matrix is built based on worldwide foreign trade data and therefore it is general for all countries.

This defines a symmetrical dissimilarity matrix of size 775×775 with elements between 0 and 1, for $0 \leq (i, j) < 775$. To join the data for the period (1998-2000), we took the average of the corresponding Φ matrices for the three years.

The proximity Π between a country c and a product p, in terms of the potential to export p in the future, depends on the proximity of the nearest exported good p' in the network (the proximity must be interpreted as a probability):

[1] Available at https://cid.econ.ucdavis.edu/data/undata/undata.html, accessed the 1st of october, 2022.



(a) Network representation of the PS for 4 diffusion cycles over the SPS and DPS.

(b) Evolution of exported products throughout diffusion cycles for different Ω.

Figure 1: Diffusion processes in the Product Space for Argentina and Germany. (a) Network representation for Argentina (top) and Germany (bottom) for the Static (left) and Dynamic (right) versions of the PS with $\Omega = 0.55$. Argentina starts with fewer products, and its productive structure shows enough potential to diffuse into several new products. Germany already produces most of the products it has potential for. The DPS presents a noticeable difference to the SPS, which is stronger for Argentina. (b) Diffusion simulations for Argentina and Germany considering different threshold values. Most scenarios converge within 2 to 4 cycles, while equilibrium values differ noticeably between SPS and DPS. Curves for $\Omega = 0.55$ correspond to the networks in panel (a).

$$\Pi^t_{c,p} = \max_{p'}\{\Phi_{p,p'} \cdot M^t_{c,p'}\} \tag{2}$$

where $\Phi_{p,p'} \in [0,1]$ are the elements of matrix Φ and $M^t_{c,p'} \in \{0,1\}$ indicates whether or not a country c features $RCA > 1$ in a product p' at time t.

Next, the diffusion process over the PS is presented. Given a threshold Ω, a country will upgrade its economy to produce all products with $\Pi > \Omega$ during the simulation cycle. We define the matrix M^t as:

$$M^t_{c,p} = \begin{cases} 1 & \text{if } \Pi^{t-1}_{c,p} > \Omega \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This matrix is initialized as follows:

$$M^0_{c,p} = \begin{cases} 1 & \text{if } RCA_{c,p} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Then, M^t is a 190×775 binary matrix which indicates if a country c develops $RCA > 1$ in a product p, while $M^0_{c,p}$ for the initial cycle ($t = 0$) comes from the data (this matrix was introduced in [6, 5]).

3 A DYNAMIC PRODUCT SPACE

Thanks to the modularity of our tool (see Section 5), extensions and modifications in the model can be introduced straightforwardly. In the original work the $\Phi_{i,j}$ proximity matrix was introduced as a static element, which we consider a debatable simplification. We therefore propose a Dynamic Product Space (DPS) model that

updates Φ by taking into account the changes in the exports of all countries as time evolves:

$$\Phi^t_{i,j} = \min\{P^t_{i,j}, P^t_{j,i}\}, \text{ with } P^t_{i,j} = \frac{\sum_c M^t_{c,i} M^t_{c,j}}{\sum_c M^t_{c,i}} \tag{5}$$

(and analogously for $P^t_{j,i}$).

The discrete time dynamics $\Phi^t = f(\Phi^{t-1})$ are a consequence of the definition of $\Pi^{t-1}_{c,p}$, see Eq. (3). We stress the difference between our dynamic approach (DPS) compared to the static version (termed here SPS) proposed in [7].

4 RESULTS

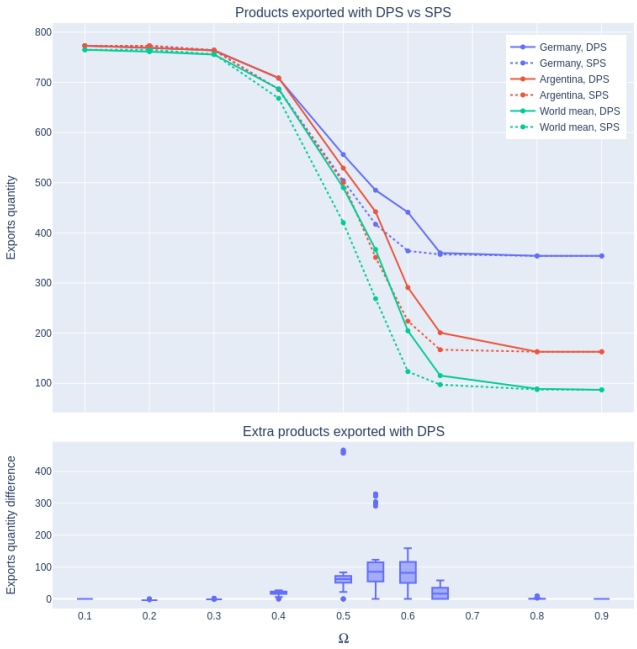


Figure 2: Comparison of products exported with SPS and DPS. For extreme values of Ω, both models behave almost identically. Yet, for intermediate values (between 0.5 and 0.6) a significant difference is observed.

In Figure 1b we compare the number of products developed, using SPS and DPS, for Argentina and Germany with different values of Ω. We can see that most simulation scenarios converge already within 2 to 4 cycles. While countries develop several new products in the first steps, from the fifth cycle onward few changes are observed.

In Figure 1a we show the diffusion process for the SPS and DPS with $\Omega = 0.55$, during 4 simulation steps. Germany starts with 354 products placed mostly in the core of the PS (cycle 0). With the SPS, Germany develops 63 new products (a 17.8% growth). Meanwhile, Argentina starts with 163 products mostly distributed in the periphery of the PS, and manages to develop up to 157 new products (a 96.3% growth) in 4 cycles. The network type of visualization can help with understanding and discovering potential development opportunities.

In Figure 2 we compare the dynamic (DPS) and static (SPS) alternatives while sweeping the threshold Ω. In the top panel we compare Argentina and Germany against the global mean, while in the bottom panel we show the distribution of the differences between SPS and DPS for all countries. Clear non trivial differences are observed between simulation scenarios for Ω between 0.5 and 0.6 suggesting that in the PS framework dynamics matter.

In some cases the advantage in exports for the global simulations are very significant, while for those with very high or very low thresholds, there does not seem to be a difference. This is consistent with the idea that, at very small values of Ω, the discovery is made very easy and at very high values it is very hard. Yet, for intermediate values of Ω between 0.5 and 0.6 we observe a significant difference. As what is being plotted is $DPS - SPS$ and the difference is positive, we interpret that the dynamic version facilitates the diffusion in these cases.

5 A MULTI-LEVEL AGENT-BASED SIMULATION FRAMEWORK

The simulation framework we developed for experimenting with the PS is based on the Emergent Behavior-DEVS (EB-DEVS) formalism [3]. It provides means to specify models mathematically that are unambiguous by construction, making them easy to understand and to reproduce. EB-DEVS is specifically designed to model complex adaptive systems, where each agent at the microscopic level is a unit of generalized behavior (with discrete and/or continuous dynamics) and, in turn, has access to information at a macroscopic level, with the capability of producing properties that emerge from the states and the interactions among agents [4].

In our PS model, each country is an agent, whereas the matrix Φ^t is a macroscopic state variable that can be influenced by all of the countries in the system. In turn, Φ^t influences the evolution of each agent (whether it will produce or not a new product). Such micro-macro interaction is formally defined in EB-DEVS.

6 CONCLUSIONS

In this work we introduced a simulation framework to test diffusion processes over the Product Space, including a dynamic extension termed Dynamic Product Space. We compared this model against its static counterpart used in [7], showing that for some relevant parameters the difference in the results can be significant. Our framework allows to flexibly study different network metrics, PS representations and types of dynamics. We expect that this tool will allow for richer simulation-based research, focused in specific countries and/or products, to explore development strategies.

REFERENCES

[1] Bela Balassa. “Trade Liberalisation and “Revealed” Comparative Advantage”. In: *The Manchester School* 33.2 (1965), pp. 99–123.

[2] Robert C Feenstra, Robert E Lipsey, Haiyan Deng, Alyson C Ma, and Hengyong Mo. *World Trade Flows: 1962-2000*. Tech. rep. 11040. National Bureau of Economic Research, 2005.

[3] Daniel Foguelman, Phillip Henning, Adelinde Uhrmacher, and Rodrigo Castro. "EB-DEVS: A formal framework for modeling and simulation of emergent behavior in dynamic complex systems". In: *Journal of Computational Science* 53 (2021), p. 101387.

[4] Daniel J Foguelman, Esteban Lanzarotti, Emanuel Ferreyra, and Rodrigo Castro. "Simulation of emergence in artificial societies: a practical model-based approach with the EB-DEVS formalism". In: (Oct. 2021). URL: https://arxiv.org/abs/2110.08170v1.

[5] César A Hidalgo. "The Dynamics of Economic Complexity and the Product Space over a 42 year period". In: *CID Working Paper Series 2009.189, Harvard University, Cambridge, MA*. 2009.

[6] César A Hidalgo and Ricardo Hausmann. "The building blocks of economic complexity". In: *Proceedings of the National Academy of Sciences* 106.26 (2009), pp. 10570–10575.

[7] César A Hidalgo, Bailey Klinger, Albert-Laszlo Barabási, and Ricardo Hausmann. "The Product Space Conditions the Development of Nations". In: *Science* 317.5837 (July 2007), pp. 482–487. ISSN: 0036-8075.

[8] John H Holland. "Complex adaptive systems". In: *Daedalus* 121.1 (1992), pp. 17–30.

Dynamics of the Ising model over highly connected random graphs with arbitray degree distribution

L. S. Ferreira[1,*] and F. L. Metz[1,2]

[1]Physics Institute, Federal University of Rio Grande do Sul, 91501-970 Porto Alegre, Brazil
[2]London Mathematical Laboratory, 18 Margravine Gardens, London W6 8RH, United Kingdom

October 25, 2022

Keywords: Dynamics, random graphs, Ising model, non equilibrium.

The theory of random graphs is of fundamental importance to describe systems where the structure of the underlying complex network affects its behaviour. One example are the Ising models, a set of discrete dynamical variables (± 1), or spins, whose state are determined by the interaction with their neighbours (specially by the coordination degree, the number of connections a spin performs), a feature completely dependent on the topology of the network at which they are defined upon. From a modelling perspective, Ising models provide a very general and simple tool with a wide range of applications, such as the study of phase transitions in statistical physics models [1], the mapping of NP problems in computer science [2] and the dynamical opinion formation on social networks [3]. In this work, we address the out-of-equilibrium dynamics of the ferromagnetic Ising model defined over highly connected simple random graphs with arbitrary degree distributions, capturing the effects of network structure on the macroscopic evolution that are not counted for when considering fully connected systems [4]. We provide the analytical solution for the discrete time evolution in the form of a closed set of dynamical equations that, for the long time limit, generalize the fixed point equations for the stationary states derived through equilibrium statistical mechanics in an early work [1]. In adittion, the generality of our formulation allows for the investigation of different noise distributions in the stochastic dynamics, that may be of interest when considering non-physical applications.

We consider a system with N Ising spins, defined upon a simple random graph whose microscopic structure is encoded by the degree sequence $(K_1, \cdots, K_N)$, where K_i is a random variable drawn from the probability distribution p_K that determines the number of connections the ith spin performs. Each connected pair of spins interacts via a constant ferromagnetic coupling energy $J > 0$. Starting from a random initial configuration, in the high connectivity limit, we show that the evolution of the macroscopic state m is given by

$$m_{t+1} = \int_0^\infty dg \nu(g) G(\beta J g u_t), \tag{1}$$

$$u_{t+1} = \int_0^\infty dg g \nu(g) G(\beta J g u_t), \tag{2}$$

where β is the inverse temperature parameter, $\nu(g)$ is the rescaled degree distribution

$$\nu(g) = \lim_{c\to\infty} \sum_{k\geq 0} p_k \delta\left(g - \frac{k}{c}\right) \tag{3}$$

and $G(x)$ is the activation function, associated with the stochastic noise distribution $\mu(\xi)$ through

$$G(x) = \int_{-x}^{x} d\xi \mu(\xi). \tag{4}$$

Regarding noise distributions, we consider the standart hyperbolic distribution μ_h, that recover thermal equilibrium results [5], and an algebraic form $\mu_a^{(\kappa)}$ that breaks detailed balance, rendering usual equilibrium statistics techniques unsuitable to the analysis of stationary states. Considering a negative binomial distribution p_K (with variance v^2), we are able to investigate the role of degree fluctuations on the system dynamics in terms of a single parameter α, associated with the rescaled distribution variance as

$$\lim_{c\to\infty} \frac{v^2}{c^2} = \frac{1}{\alpha}. \tag{5}$$

Therefore, in the limit of $\alpha \to \infty$ we recover the behaviour of a fully connected model, where the usual mean field theories fit perfectly, while the effects of degree fluctuations arises for finite α. We characterize the relaxation time as a function of α for different noise distributions, obtaining for the critical relaxation ($t >> 1$)

$$m_t^h \sim t^{-\frac{1}{2}}, \quad m_t^a \sim t^{-\frac{1}{2\kappa}}. \tag{6}$$

These results are presented on the left panel of figure (1). In the limit of large t (when $u_{t+1} = u_t, \forall t$), the set (1) and (2) is reduced to a fixed point equation for u that defines m, whose solution is presented in the right panel of figure (1). We can see the tendency towards the fully connected Ising ferromagnet (or the Currie-Weiss model, in the case of hyperbolic tangent activation [6]) as α increases. We also obtain the critical temperature T_c as a function of α, reassuring the result from [1], and go further to the critical exponents of the magnetization, obtaining for $T \to T_c$

$$m_h \sim C_h \left(\frac{T - T_c}{T_c}\right)^{\frac{1}{2}}, \quad m_a \sim C_a(\kappa) \left(\frac{T - T_c}{T_c}\right)^{\frac{1}{2\kappa}}. \tag{7}$$

Remarkably, the values of the critical equilibrium exponents are the same of the critical dynamical ones.

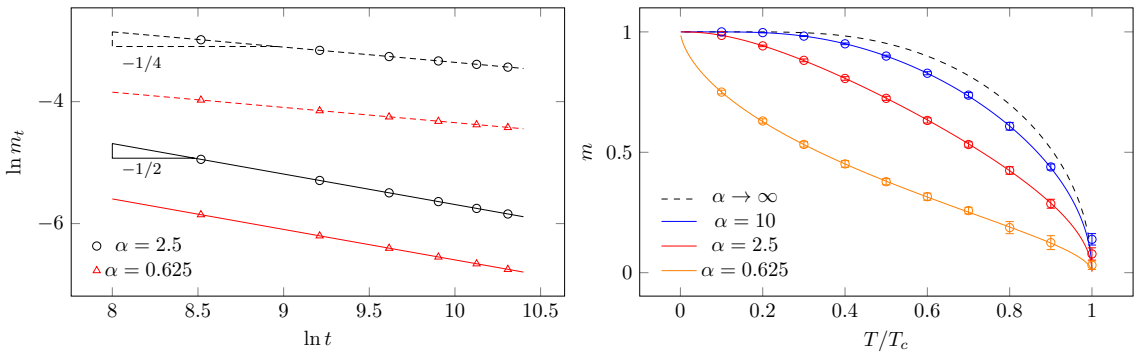


Fig. 1: The left panel presents the logarithmic relaxation for the algebraic distribution with index value $\kappa = 1$ (solid lines) and $\kappa = 2$ (dashed lines), with slope given by $-\frac{1}{2\kappa}$; the symbols are results of iteration of (1) and (2). The right panel presents a comparison between stationary limit of (1) and (2) (lines) and Monte Carlo simulations (symbols). Black dashed line denote the fully connected regime.

Overall, our work introduces a family of Ising models over random graphs that retain the effect of both topological structure and noise distribution whose non-equilibrium dynamics can be solved exactly, presenting insights on the network effect on the dynamics.

References

[1] F. L. Metz and T. Peron. Mean-field theory of vector spin models on networks with arbitrary degree distributions. *Journal of Physics: Complexity*, 3(1):015008, 2022.

[2] Andrew Lucas. Ising formulations of many np problems. *Frontiers in Physics*, 2, 2014.

[3] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, May 2009.

[4] A. C. C. Coolen. Statistical mechanics of recurrent neural networks ii. dynamics, 2000.

[5] A. C. C. Coolen. Statistical mechanics of recurrent neural networks i. statics, 2000.

[6] M Kochmański, T Paszkiewicz, and S Wolski. Curie–weiss magnet—a simple model of phase transition. *European Journal of Physics*, 34(6):1555–1573, oct 2013.

Effects of mobility-based dependency networks on economic resilience

Takahiro Yabe[1], Bernardo Garcia Bulle Bueno[1], Morgan Frank[1,2],
Alex 'Sandy' Pentland[1], Esteban Moro[1,3]

[1]MIT, [2]University of Pittsburgh, [3]Universidad Carlos III de Madrid

Keywords: economic resilience, human mobility, urban networks

Introduction

Quantifying the economic costs of businesses caused by extreme shocks, such as the COVID-19 pandemic and natural disasters, is crucial for developing preparation, mitigation, and recovery plans. Conventionally, survey data have been the primary source of information used to measure losses inflicted on businesses by negative shocks, however, drops in foot traffic quantified using large scale human mobility data (e.g., mobile phone GPS) have recently been used as low-cost and scalable proxies for such losses, especially for businesses that rely on physical visits to stores, such as restaurants and cafes (Yabe et al., 2019). Such studies and analyses often quantify the losses in foot traffic based on individual points-of-interest (POIs), neglecting the interdependent relationships that may exist between businesses and other facilities. For example, university campus lockdowns imposed during the COVID-19 pandemic may severely impact foot traffic to student-dependent local businesses. Such dependent relationships between businesses could cause secondary and tertiary cascading impacts of shocks and policies, posing a significant threat to the economic resilience of business networks (Zhai and Yue, 2021).

Data and Methods

To identify such cascading effects, we build a "dependence demand network" of business using mobility data. We used large-scale anonymous, privacy-enhanced mobility data of more than 200K devices from the Metropolitan Boston Area collected in 2019 to 2021 (Moro et al., 2021). The dependency scores were computed based on the foot traffic patterns before the pandemic (September 2019 to January 2020). We compute the dependence of a target POI i on a source POI j by $dep(i,j) = |s_i \cap s_j|/|s_i|$, where s_i and s_j denote the sets of users who visit POIs i and j respectively. Because the denominator is based on the number of users who visit the target POI i, $dep(i,j) \neq dep(j,i)$. This is a simple but intuitive measure that considers the asymmetric nature of dependencies between POIs. The set of users who visit each POI in a

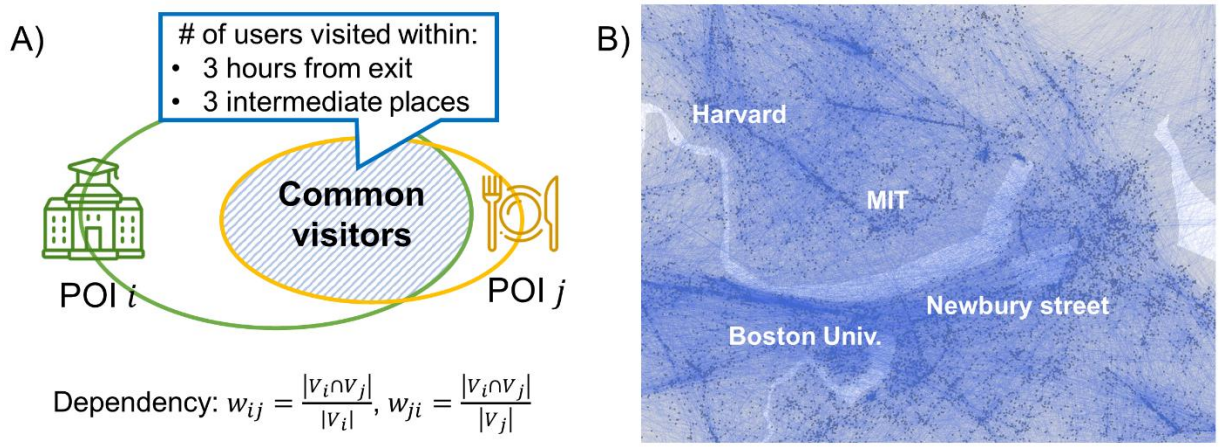


Figure 1. Mobility based dependency network. A) Definition of mobility-based dependency network. Dependency between POIs i and j are computed using the number of common and total visitors to each POI. B) Dependency network of POIs in Boston metropolitan area shows high clustering in local districts such as Harvard, MIT, Boston Universities and Newbury Street.

specific period is computed using mobility data collected from mobile phone devices. Figure 1A shows an overview of the methods on how the dependency between two POIs is computed. Figure 1B shows the dependency network in the Boston metropolitan area. Commercial and college districts, such as MIT, Harvard, and Boston University, as well as Newbury Street, can be seen in the network.

Network Characteristics

To quantitatively understand the characteristics of the dependency network, we constructed several null network models and compared various network statistics. For this analysis, edges with dependency weights greater than 0.05 were labeled as edges and the resulting unweighted directed network was analyzed. Null network models were created by randomly shuffling the links from the original (real) network in different ways. The most constrained random network controls the in- and out-degrees of each node and also keeps the distance distribution of the edges ('spatial configuration model'). As shown in Figure 2, network metrics such as the clustering coefficient, reciprocity, and transitivity are significantly and substantially higher in the real network compared to all null networks including the spatial configuration model, which indicates that POIs in the real network are more locally clustered and dependent on each other.

Network	Clustering coefficient	Reciprocity (% edges returned)	Transitivity (% triangles closed)
Real Network	**0.054**	**0.131**	**0.030**
Erdos-Renyi	< 0.001	< 0.001	< 0.001
Spatial E-R	0.011	0.020	0.021
Configuration	< 0.001	< 0.001	< 0.001
Spatial configuration (null model)	**0.003**	**0.004**	**0.002**

Figure 2. Network statistics of the real dependency network compared with multiple null network models, including the spatial configuration model which controls for the in- and out-degrees of each node and also the overall distance distribution of the edges. Analysis shows that metric related to local clustering, such as the clustering coefficient, reciprocity, and transitivity are all significantly greater in the real network compared to the null models.

Economic Impacts of Dependency Networks

To measure how such dependency relationships may affect the resilience of businesses to external shocks, we analyze how dependency on visits from office POIs and university/college POIs affects the magnitude of disruption that restaurant and café POIs experienced during the COVID-19 pandemic. The dependency weights were computed based on the foot traffic patterns before the pandemic, and as the objective variable, the percentage drop in the number of foot traffic to POIs were used. The reduction in foot traffic to restaurant and café POIs were calculated relative to the foot traffic in January to February 2020. First, the intuitive relationship between dependency on colleges of restaurants were investigated. Figure 3A shows the relationship between the dependency on visits from/to offices and university/college POIs (x-axis, between 0 and 1) and the foot traffic levels compared to pre-pandemic levels (y-axis, in %). The relationships are shown for two periods (April 2020 and April 2021). We observe a negative trend between

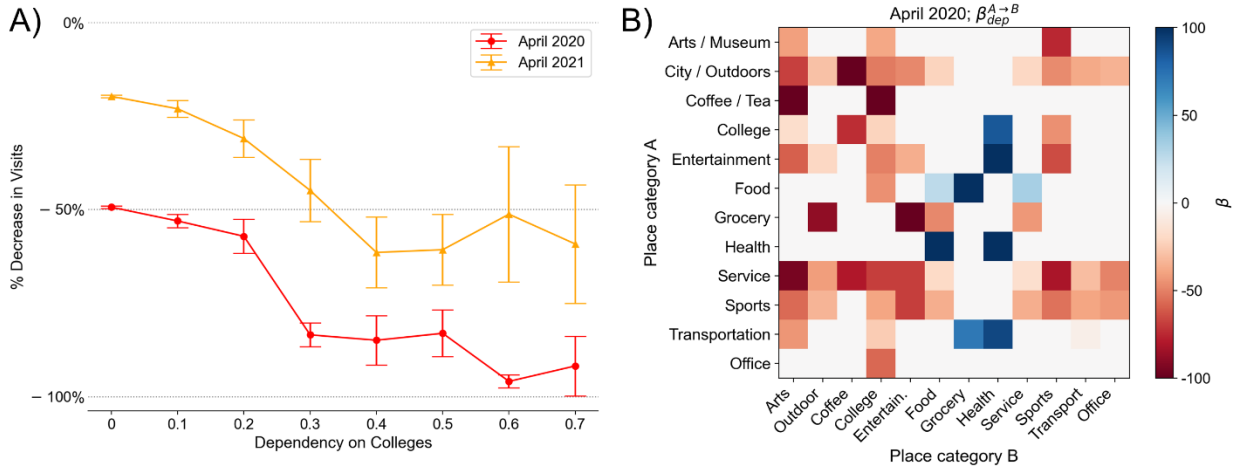


Figure 3. Economic impacts of dependency networks. A) Food and coffee places that were more dependent on colleges suffered larger losses in foot traffic during the pandemic. B) POIs that were dependent more on many of the POI categories suffered more, however, food and health places that were dependent more on grocery and health facilities performed better than other POIs.

dependency and foot traffic for both periods, indicating that restaurants and cafés that were more dependent on offices and universities experienced more substantial negative impacts of the non-pharmaceutical intervention policies during COVID-19. Similarly, Figure 3B shows the regression coefficients of dependency between all POI category pairs. A negative coefficient β indicates that the higher the dependency on place category B, POIs in place category A performed worse during the pandemic. On the other hand, relationships in blue color indicate that POIs that were dependent on grocery and health places performed better than those which did not. This methodology enables us to further investigate the effects of mobility-based dependency networks, and its impacts on the local economy with various hypothetical urban shocks, including not just pandemics but also natural disasters (Sadri et al., 2018), public transport disruptions, and also positive interventions such as festivals and public events.

Acknowledgment

We thank Spectus for providing the mobility data used for the analysis through their Social Impact program.

References

Yabe, T., Zhang, Y., & Ukkusuri, S. V. (2020). Quantifying the economic impact of disasters on businesses using human mobility data: a Bayesian causal inference approach. *EPJ Data Science*, *9*(1), 36.

Zhai, W., & Yue, H. (2022). Economic resilience during COVID-19: An insight from permanent business closures. *Environment and Planning A: Economy and Space*, *54*(2), 219-221.

Moro, E., Calacci, D., Dong, X., & Pentland, A. (2021). Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications*, *12*(1), 1-10.

Sadri, A. M., Ukkusuri, S. V., Lee, S., Clawson, R., Aldrich, D., Nelson, M. S., & Kelly, D. (2018). The role of social capital, personal networks, and emergency responders in post-disaster recovery and resilience: a study of rural communities in Indiana. *Natural hazards*, *90*(3), 1377-1406.

Ergodic sets in directed networks: a dynamics-based simplification

Erik Hörmann Renaud Lambiotte

November 7, 2022

Keywords: ergodic sets, directed networks, random walks,coarse-graining

Abstract

We introduce ergodic sets, strongly connected components of directed networks [1] that are connected in either direction to the rest of the graph, but not in both directions. We use these newly introduced structures to propose a coarse-graining method for large complex directed networks, which preserve the random walk dynamics of the original network.

1 Introduction

We start by formulating a mathematical definition of the generalized ergodic sets, from a purely structural construction that only depends on the network wiring.

Definition 1. *Let $\mathcal{G} = (V, E)$ be a directed graph. We say $X \subseteq V$ is an **ergodic set** if X is strongly connected and **either** of the two following sets of conditions hold:*

1. *$X = V$*
2. *$X \subset V$ and **all** the following conditions hold:*
   (a) *if $\exists x^* \in X$, $v^* \in V \setminus X$ such that $(x^* \to v^*) \in E$, then*
   $$\forall x \in X, \forall v \in V \setminus X \Rightarrow (v \to x) \notin E \tag{1}$$
   (b) *if $\exists x^* \in X$, $v^* \in V \setminus X$ such that $(v^* \to x^*) \in E$, then*
   $$\forall x \in X, \forall v \in V \setminus X \Rightarrow (x \to v) \notin E \tag{2}$$

Furthermore, we call input an ergodic set that has out-edges to the rest graph (i.e. one for which equation (2) is non-empty), and an output an ergodic set that has in-edges from the rest graph (i.e. one for which equation (1) is non-empty). Note that an ergodic set must either be an output, or an input, or a strongly connected component disjoint from the rest of the graph. Furthermore, an ergodic set X cannot be both an input and an output, as the two conditions are mutually contradicting.

2 Coarse-graining with ergodic sets

The ergodic sets has been purposely defined so that they can be used to simplify the structure of complex directed networks. One of the issues of directed networks are the presence of sources and wells, nodes that have zero in- and out-degree, respectively. These structures have risen to prominence in the seminal paper of Page and Brin [2], where these types of nodes precluded web crawler to explore the entire word wide web network, and this limitation prompted the development of the well-known PageRank algorithm by the same authors. The concept of ergodic sets allows us to generalise sources and wells to groups of more than one node.

2.1 Detecting ergodic sets

The first step is to find the strongly connected components of the network. To do so, we rely on the algorithm implemented in the standard library. For example, the NetworkX algorithm uses Tarjan's algorithm [3] with Nuutila's [4] modifications.

We now need to implement an algorithm to efficiently check whether the strongly connected components satisfy the additional conditions in the definition of ergodic sets, equations (1) and (2).

We use the sum of the in- and out-degrees of all nodes in the ergodic set $v \in X \subseteq V$,

we compare these two quantities in the subgraph X properly and in X as induced subgraph X_I. More formally, the following relations hold:

$$\left(\sum_{v \in X} d_{out}(v) - \sum_{v \in X_I} d_{out}(v) \neq 0 \Rightarrow \sum_{v \in X} d_{in}(v) - \sum_{v \in X_I} d_{in}(v) = 0 \right) \Leftrightarrow \text{equation (1)} \tag{3}$$

$$\left(\sum_{v \in X} d_{in}(v) - \sum_{v \in X_I} d_{in}(v) \neq 0 \Rightarrow \sum_{v \in X} d_{out}(v) - \sum_{v \in X_I} d_{out}(v) = 0 \right) \Leftrightarrow \text{equation (2)} \tag{4}$$

which allows us to quickly and efficiently test whether the strong connected component is also an ergodic set.

2.2 Coarse-graining

If X is an output, just collapse all vertices in one and keep all the in-edges. If a single vertex from $v \in V \setminus X$ points to more than one node of a single ergodic set, only add one edge. If a single vertex from $v \in V \setminus X$ points to multiple vertices across different outputs, $O_1, O_2, \ldots, O_n$ add one edge per ergodic set with weight proportional to the number of vertices in the ergodic set to which $v$is connected to.

If X is an input, proceed as follows. Substitute X with a single node v_X and connected with all the nodes w for which $\exists v \in X : (v, w) \in E$. all the out-and in-degrees from the other vertices of the graph an in- or out-edge with weight

$$\omega_w = \frac{1}{\sum_{v \in X} d_{in}(v)} \sum_{v \in X} \frac{d_{in}(v)}{d_{out}(v)} \mathbb{I}_E(v, w) \tag{5}$$

in which

$$\mathbb{I}_E(v, w) = \begin{cases} 1 & \text{if} (v, w) \in E \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

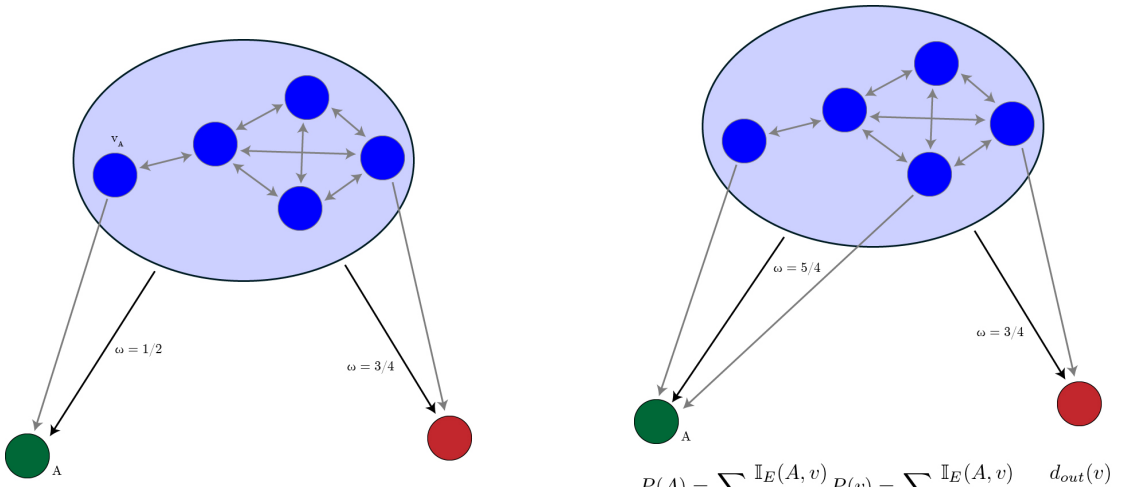
It is clear by construction that the reachable vertices w, as well as the probability of reaching them from the ergodic set X does not change when substituting X with its coarse-grained vertex v_X.

In order to have an equivalent dynamics, however, we need to set the hitting time for the outgoing nodes such that is the same before and after the substitution. We have one more degree of freedom to fix this extra requirement, that is, the weight ω_s of a self loop for v_X. Using the fact that the exit time of a random walk is an exponential variable, we see that we need to set ω_s such that:

$$\frac{\omega_s}{\omega_s + \sum_i \omega_i} = \sum_{v \in X} \frac{d_{in}(v)}{\sum_{w \in X} d_{in}(w)} \frac{d^{\dagger}_{out}(v)}{d_{out}(v)} \tag{7}$$

in which we have indicated the (out-)degrees of the vertices in X as a proper subset of V with $d(v)$, while $d^{\dagger}(v)$ indicates the (out-)degree of v as a vertex of X_I, the subgraph induced by X.

Finally, a quick note on the quantities involved. While the notation is not the most user-friendly, all the quantities involved can be easily calculated in linear time using the appropriate data structure for the graph (as they are just degrees) and so there is no major barrier to apply the method for large graphs.



$$P(A) = \frac{1}{d_{out}(v_A)} P(v_A) = \frac{1}{d_{out}(v_A)} \frac{d_{in}(v_A)}{\sum_{v \in X} d_{in}(v)} \propto \omega_A$$

$$P(A) = \sum_{v \in X} \frac{\mathbb{I}_E(A, v)}{d_{out}(v)} P(v) = \sum_{v \in X} \frac{\mathbb{I}_E(A, v)}{d_{out}(v)} \frac{d_{out}(v)}{\sum_{w \in X} d_{in}(w)}$$
$$= \frac{1}{\sum_{w \in X} d_{in}(w)} \sum_{v \in X} \frac{d_{in}(v)}{d_{out}(v)} \mathbb{I}(A, v)$$

Figure 1: Graphical representation of the weight assignment rules for the output sets

3 Perspectives

Ergodic set are a simple yet powerful tool that allows to simplify the structure of directed networks using only the underlying structure of the network. They could be deployed in the analysis of real world network, and particularly food chains, which naturally presents clusters of species that are either apex predator or at the bottom of the food chain.

They also can further used to measure the directedness of networks, which can be defined from a random walker $(X_n)_{n \geq n}$. We can use the distribution of the hitting times of the outputs, conditioned on the walker starting in an input, to infer how "directed" the network is. This is the same idea as the hodge decomposition, which we plan to use as a reference tool to test the measure developed as described.

Finally, we can also use the construction in this work to form an insight on how much the network is mixing, by defining the *ergodic mixing matrix*. Take a random walker $(X_n)_{n \geq n}$ and define:

$$C_{ij} = \mathbb{P}\left(\lim_{t \to +\infty} X_t \in \mathcal{O}_j | X_0 \in \mathcal{I}_i \right) \tag{8}$$

in which $\{\mathcal{O}_1, \ldots, \mathcal{O}_n\}$ is the set of all output, and $\{\mathcal{I}_1, \ldots, \mathcal{I}_m\}$ the set of all outputs. This can be used to reduce the entire network core and obtain an extreme coarse-grained representation of the network, thus showing a remarkable flexibility of the construction, which can coarse-grain network at different levels.

References

[1] Newman M., *Networks*,Oxford university press (2018).

[2] Brin S., Page L. *The anatomy of a large-scale hypertextual web search engine* Computer networks and ISDN systems, (1998)

[3] Tarjan R., *Depth-first search and linear graph algorithms* SIAM Journal of Computing, (1972)

[4] Nuutila E., Soisalon-Soinen E., *On finding the strongly connected components in a directed graph* Information Processing Letters, (1994)

Exploring Informative Scales of Labor Networks in Argentina

Sergio A. De Raco*, Viktoriya Semeshenko†, and Juan I. Vázquez Broquá‡

Keywords— labor mobility, complex networks, granularity, effective information

Labor flows are a key factor explaining economic activity through the interplay of workers' supply and firms' employment demand in the labor market. Particularly, job-to-job transitions are relevant labor flows, with recognized pro-cyclical behavior [6] that carry tacit information about the relevance of past jobs' experience for new employers, specially those occurring between firms with different economic activities. Traditionally, economists analyze labor flows with data at high level of aggregation of the standard classifications of productive activities, in order to correlate it with conventional national accounts data of sectoral activity.

De Raco and Semeshenko [1, 2] studied inter-industry labor flows in Argentina from administrative data at high level of details (four digits of the ISIC[1] Rev.3 classification) using a network representation between economic sectors, and revealed that labor networks extracted are typically very dense, not sparse, with clear core-periphery structures, and present small-world properties.Although these microscale networks provide new and useful information, they also pose several challenges for their interpretation and applications in, for example, policy design and analysis.

Ensuing, Semeshenko and De Raco [3] inspect the evolution of the connectivity structure in (binary) labor flows networks at different scales of aggregation using standard classification aggregation schemes, and evidenced that more disaggregated data can bring insights into the evolution of the connectivity of the network. Data granularity defines the size of a network, which in turn determines its structure and functioning. Naturally, several questions and issues arise related to the granularity of the data. What is an adequate level of granularity providing enough information of sectoral aggregation that allows to characterize employment flows as well as the structure of inter-industrial relations? When is best to look for more detail in labor flows? What is the best informative scale of the network?

Finding an appropriate scale/size for a network can be made with different algorithmic techniques *via* two alternative schemes [4]: (a) by changing the observation scale, coarse-graining or grouping their nodes by a common attribute; or, (b) by keeping the observation scale through filtering or pruning edges by keeping only those meeting some specified criteria. In this work we search for an appropriate and informative scale of labor flows networks using the efficient entropy method [5]. The network is reduced using a measure called effective information which

*Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. CONICET-Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina

†Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina

‡Universidad Catolica Argentina. Universidad de Buenos Aires

[1]The International Standard Industrial Classification of All Economic Activities (ISIC)

is a general measure for causal interactions because it uses perturbations to capture effectiveness of the mechanisms of the system in relation to the size of its state space. Networks with higher effective information contain more information in the ties between the nodes.

Data. We use administrative data of year-to-year private labor flows between economic activities at four digits (ISIC Rev.3 classification) between 1996 and 2020, provided by the Ministry of Labor, Employment and Social Security. After filtering out temporary employment agencies and other (13) economic activities for which there is no employment data available for the whole period, we use the remaining 287 economic activities, or sectors, for the analysis. Regarding labor flows, we focus on intersectoral labor flows that is to say we leave out the analysis of transitions occurring within the same sector. Using sectors as nodes and labor flows as edges, we build 24 interanual labor flows directed and weighted node-aligned networks.

Methods. Network analysis is typically performed on the full microscale representation of a network, and it can be extremely noisy and uninformative. Networks can have macroscales that can exhibit different network properties than their underlying microscales. The procedure to find informative macronodes is similar to community detection, is focused on subgraphs that have more in-group connectivity that out-group. The connectivity of a network contains information about the interactions between the nodes (sectors), their associations and dependencies. We use here the global network measure -effective information (entropy-based)- which measures the uncertainty contained in paths along nodes and links, and best captures the connectivity information of the network [5]. Using this technique, a network can be recast into a new one, wherein subgraphs of nodes are grouped into individual macronodes, reducing the size of the original network. These macronodes summarize the behavior of the subgraph in a manner that it increases the effective information in connectivity (causal emergence) of the network, when compared to the original network.

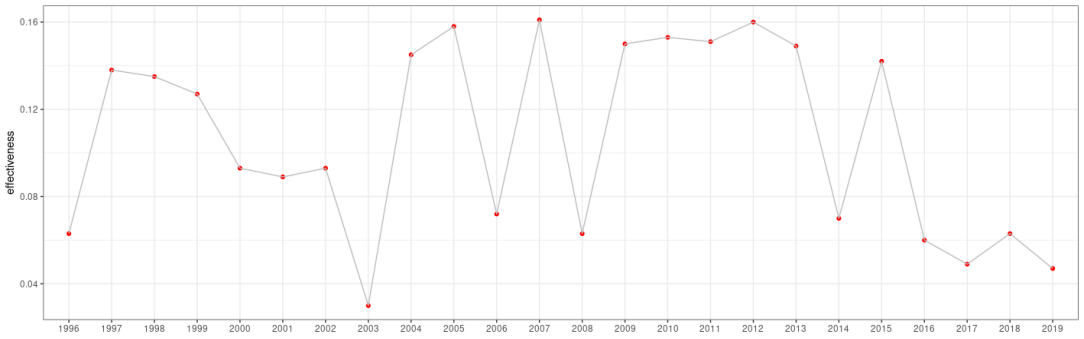


Figure 1: Effectiveness of the network 1996-2020. High effectiveness observed in periods of relative macroeconomic stability, low effectiveness in periods of crisis or lower economic activity.

In preliminary experiments, the existence of macronodes increases the informative quality of the system and confirms the causal emergence measured as effectiveness in the periods under study. The obtained effectiveness exhibits bounded volatility over time during the periods analyzed in the range $[0.03, 0.17]$, with high effectiveness in years of relative macroeconomic stability, and low values in years of macroeconomic crisis or lower economic activity, see Fig. 1. Additionally, the existence of a significant number of micronodes that share macronodes over time was confirmed, raising expectations of the partial stability of some mesoscale organization of the economic sectors involved.

References

[1] De Raco, S., Semeshenko, V.: Labor mobility and industrial space in Argentina. Journal of Dynamics & Games, **6**(2), pp. 107–118 (2019)

[2] De Raco, S., Semeshenko, V.: The network structure of inter-industry labor mobility in Argentina. In: 6th Conf. Regulating for Decent Work, ILO, Geneve, July (2019)

[3] Semeshenko, V., De Raco, S.: Analysis of the evolution of labor market flows in Argentina. In: AGRANDA 2021, Simposio Argentino de Ciencia de Datos y Grandes Datos (JAIIO), pgs. 20-24 (2021)

[4] Serrano, M. Á., Boguná, M., Vespignani, A.: Extracting the multiscale backbone of complex weighted networks. Proceedings of the national academy of sciences, 106(16), 6483–6488 (2009)

[5] Klein, B., Hoel, E.: The emergence of informative higher scales in complex networks. Complexity, (2020)

[6] Mukoyama, T.: The cyclicality of job-to-job transitions and its implications for aggregate productivity. Journal of Economic Dynamics and Control, 39, 1-17 (2014)

Framework for developing quantitative agent-based models based on qualitative expert knowledge: *an organised crime use-case*

Frederike Oetker, MSc.

In order to model criminal networks for law enforcement purposes, a limited supply of data needs to be translated into validated agent-based models [1, 2]. What is missing in current criminological modelling is a systematic and transparent framework for modelers and domain experts that establishes a modelling procedure for computational criminal modelling that includes translating qualitative data into quantitative rules [3, 4]. For this, we propose FREIDA (Framework for Expert-Informed Data-driven Agent-based models).

Throughout the paper, the criminal cocaine replacement model (CCRM) will be used as an example case to demonstrate the FREIDA methodology.

For the CCRM, a criminal cocaine network in the Netherlands is being modelled where the kingpin node is being removed, the goal being for the remaining agents to reorganize after the disruption and return the network into a stable state. The agents are simultaneously embedded in multiple social and business networks and possess heterogeneous individual attributes which determine the probability of shared ties, and the possibility of new relations being formed. Qualitative data sources such as case files, literature and interviews can be translated into empirical laws, and combined with the quantitative sources such as databases form the three dimensions (environment, agents, behaviour) of a networked ABM.

Finally, FREIDA introduces sensitivity statements and validation statements to transition to the computational model and application phase respectively. In the last phase, iterative sensitivity analysis, uncertainty quantification and scenario testing eventually lead to a robust model that can help law enforcement plan their intervention strategies.

Keywords: methodological framework, criminological modelling, computational networks, validation methods

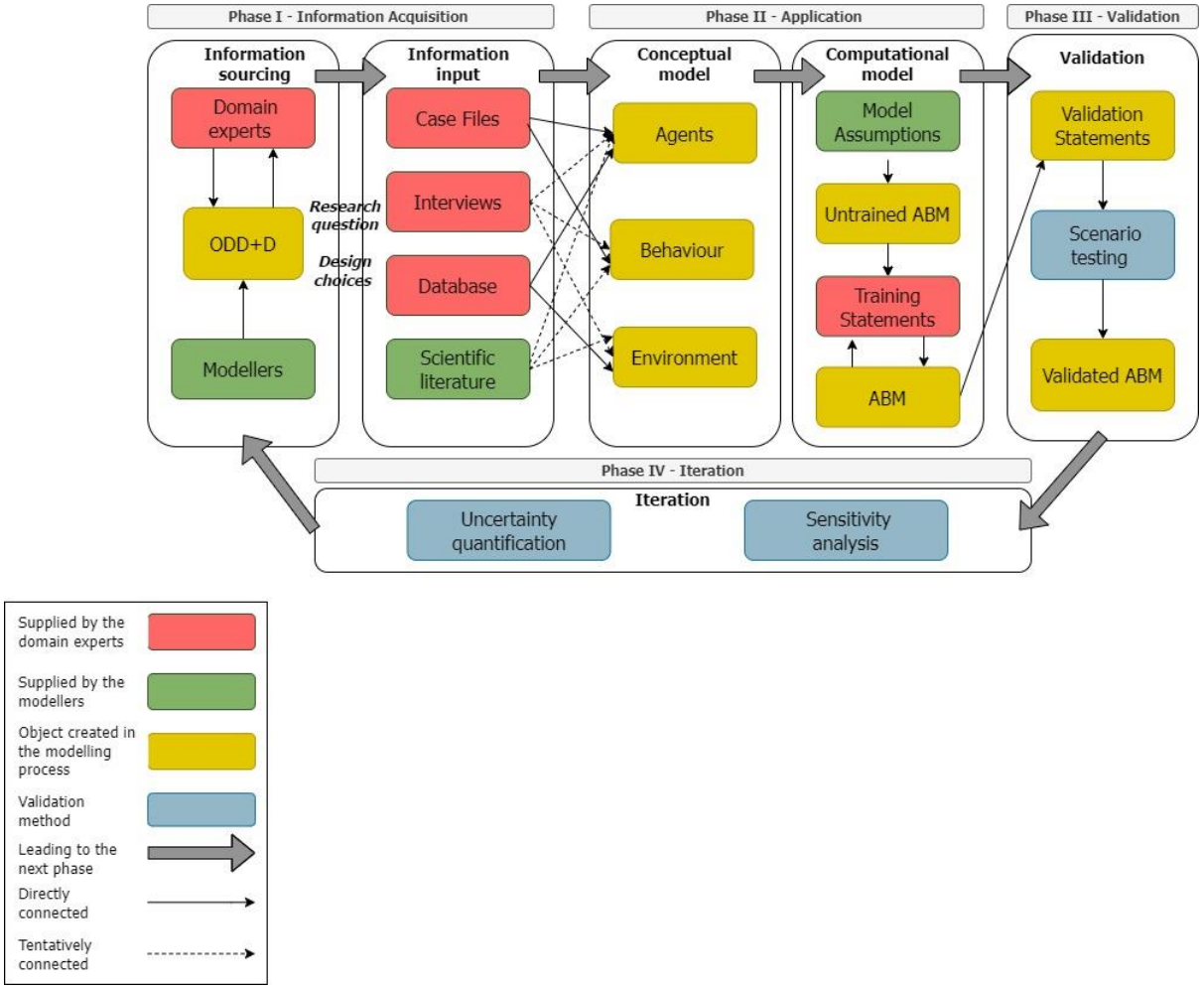


Figure 1: FREIDA methodology, with the four methodology phases (information acquisition, application, validation and iteration).

[1] Roks, Robert & Bisschop, Lieselot & Staring, Richard. (2021). Getting a foot in the door. Spaces of cocaine trafficking in the Port of Rotterdam. Trends in Organized Crime. 24. 10.1007/s12117-020-09394-8
[2] Rosés Brüngger, Raquel & Kadar, Cristina & Pletikosa, Irena. (2016). Design of an Agent-Based Model to Predict Crime (WIP)
[3] Müller et al., Describing human decisions in agent-based models - ODD+D, an extension of the ODD protocol, 2013
[4] Transparent and Comprehensive model Evaluation) (Towards better modeling and decision support: Documenting model development, testing, and analysis using TRACE, 2014

From networks to flows: Using flow maps to understand mobility patterns in cattle trade

Sima Farokhnejad[1], Eraldo Ribeiro[2] and Ronaldo Menezes[1,*]
[1]Department of Computer Science, University of Exeter, UK
[2]Department of Computer Science, Florida Institute of Technology, USA
sf503@exeter.ac.uk, eribeiro@fit.edu, *r.menezes@exeter.ac.uk

Keywords: cattle trade networks, flow maps, cattle movement

The understanding of mobility patterns is very important in a wide range of research areas [1, 2], and it has a direct impact on policymaking in urban environments, disease spreading analyses, and control of epidemics [3, 4], to name a few. Despite the power of networks in modelling mobility, alternative methods may be more useful or complementary in capturing idiosyncrasies of mobility patterns, depending on the datasets, or areas being investigated. Mobility datasets are often large and usually contain plenty of individual records. This feature adds too many details to the network generated from the dataset, making the modelling itself often confusing; simulation becomes time-consuming, and it is difficult to have a global analysis of mobility. Furthermore, to analyse global patterns of mobility during a period of time, it is critical to be able to look beyond microscopic details and able to see global patterns. To add to these challenges, mobility datasets are always subject to uncertainty [5]. Many datasets, including human mobility and livestock trade, show evidence of uncertainty. As a result of the forgotten or unreported records, or the scale of the datasets, the networks generated from the dataset are anything but flawless, which could affect traceability or reliability of the results.

The dataset of cattle movement in Brazil is an example of scale and uncertainty; the state we deal in this work (Minas Gerais) produces more cattle than many countries around the world. Economic pressures in poorer nations coupled with poor tracking infrastructure make it easier for people to trade cattle without reporting it. According to a recent report in a political magazine [1], these unreported trades occur frequently in Brazil as a mechanism akin to money laundering. The Amazon's vast national forest, which cannot be cultivated, has been ravaged over the last two decades by miners and thieves of public lands looking for hardwoods and new pastures for cattle raising. Animal Transit Guides (ATGs) indicate that there are over 91,000 cattle heads raised on stolen land, leading to a number of cases where cattle are transported clandestinely, without registration in any ATG.

As a result, there is missing information in these datasets. Due to its imperfection, the dataset loses the ability to trace the path of trades at some point. However, regardless of these microscopic details, analysing the path of the trades is extremely relevant, as well as recognising the global mobility patterns throughout time. The use of flow maps as an alternative representation of trade is a novel method for understanding trade patterns over time. It is a general approach that could be applied to any type of mobility dataset, but here we focus on cattle movement. Flow maps are constructed using all the details of individual trades. However, they do not need to include details in each trade to get to a model of global behaviour, which makes it much more compatible for the use in large and complex mobility datasets.

The mobility dataset is used to generate a vector field (on a map) in order to produce a flow map. There are so-called critical points that need to be analysed. Observing the vector field behaviour near these points gives us insights into its characteristics. It is our opinion that *sinks* and *sources* play a major role in epidemiological research as destinations and starting points of infection, respectively. An analysis of sink-source dynamics in mobility trades is a valuable approach that is captured in moving from network methods to flow methods. In order to understand the global behaviour of the trade patterns, it is essential to analyse the dynamics of the critical points on the flow maps of trades of different months. Monitoring the changes in sinks and sources on maps within time windows (monthly, for instance) can be effective in finding the level of similarity between trade patterns in two consecutive months (or other periods of time).

[1] `https://piaui.folha.uol.com.br/materia/lavagem-da-boiada/` *The Laundering of Cattle (In Portuguese)* (last accessed: 1 Nov 2022).

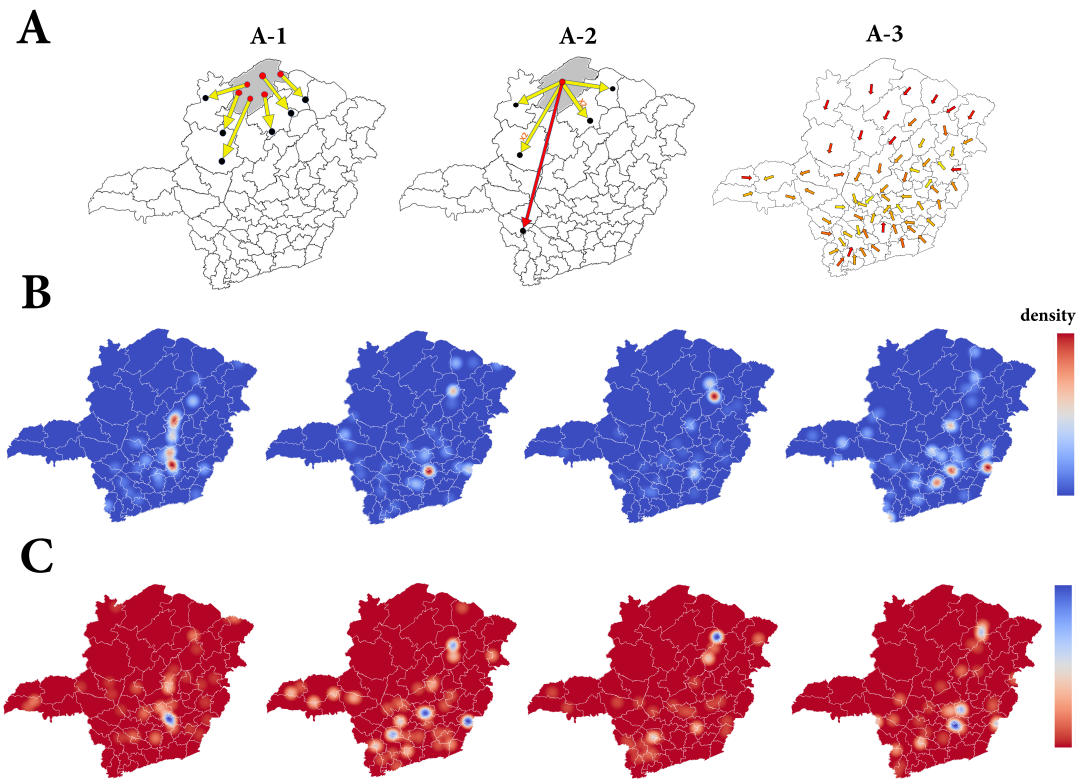


Figure 1: **A) An illustration of how to generate a flow map from a cattle mobility dataset. A-1**) Take the grey part as an example and concentrate on the trades happening between it and other parts. **A-2**) The trades are captured with a vector starting from the centre of the location (in grey) and ending at the centre of the destinations (yellow vectors). All vectors are then aggregated using a relevant method. Here, we add the vectors into a final vector (red vector), which we assign to the grey location. **A-3**) Visualisation of what we call the flow map. For a better visualisation, all vectors are displayed with the same length. The colour of each vector corresponds to a range that shows the size of the vector. **B) Sinks and C) Sources are visualised in flow maps for four consecutive months.** Initial flow maps are generated based on trades happening within specific months. A triangle-based interpolation method is used to create a vector field from the initial flow map. We found critical points in this vector field and identified sinks and sources. Sink and source areas are shown on monthly trade flow maps. In **B**, red areas indicate locations with a high density of sinks in the vector fields. Dark blue shows the areas with zero density of sink points. In **C**, the emphasis is on sources. Dark blue areas indicate high-density of sources, and dark red areas are the ones empty of source points.

In this work, we examine a dataset related to cattle movement in the state of Minas Gerais in Brazil. Figure 1-A is a toy example of how trades can be used to produce an output vector for each part (here each part is a micro-region in the state, although the regions be for a higher or lower granularity if needed). Trades within one part are first removed from the records; we do not consider trades internal to one area. As an example, consider the grey part in Figure 1-A-1, each trade originating from this part is captured as a vector. In Figure 1-A-2 for each trade between a location in the grey part and a destination in another part, a vector with the start point in the centre of the grey part and the end point in the centre of a destination part is assigned. By employing a method for the combination of vectors (here the average of the vectors is determined), the vectors of the grey part are merged into a final vector (red vector in Figure 1-A-2).

The result is a flow map for the trades that happen over a given period of time. Figure 1-A-3 shows a flow map with vectors in the centres of micro-regions, which could be regarded as a scatter vector field. Having an approximation of the vector value at each point of the map allows us to easily analyse the global pattern of the vector field. The interpolation methods involve constructing vectors for new points based on known scatter vectors. A triangle-based interpolation [6] is the technique we used to estimate the value of vectors at points

inside the map that have not been calculated.

Points with vector sizes equal to zero are critical points in a vector field. Classification of these critical points as sink, source, and saddle points is according to the sign of the eigenvalues of the Jacobian matrix [7]. Understanding how sink and source patterns change from month to month can help us understand the level of dynamics we have in the phenomenon being captured and help policymakers reach better decisions. Similarities or differences in the patterns of sinks and sources in different months of the year determine how predictable a year is in terms of mobility. Heatmaps in Figure 1-B and 1-C show the density of sinks and sources in flow maps generated from monthly trades. Over a period of four consecutive months, the sinks and sources area are recognised.

We developed an approach that can be applied to any type of mobility dataset which contains sources and destinations in trajectories. It provides an alternative/complementary tool to network methods for analysing the dynamic patterns of mobility, and it is independent of intermediate steps reached between source and destination of moving entities. Availability of numerous techniques to combine trades originated from each part to produce the final vector for that part is a strong aspect of this approach. As a result, it becomes more relevant to a particular type of mobility dataset and analysis goals; for instance, one could consider the number of cattle heads being transported as a factor in the value of the vectors. Additionally, any kind of dataset that contains origin-destination values can be converted to a flow map using this approach. The dynamic behaviour of the network during the time could then be evaluated using flow maps.

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[2] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.

[3] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009.

[4] A. Arenas, W. Cota, J. Gómez-Gardeñes, S. Gómez, C. Granell, J. T. Matamalas, D. Soriano-Paños, and B. Steinegger, "Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions," *Physical Review X*, vol. 10, no. 4, p. 041055, 2020.

[5] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–41, 2015.

[6] D. Watson and G. Philip, "Triangle based interpolation," *Journal of the International Association for Mathematical Geology*, vol. 16, no. 8, pp. 779–795, 1984.

[7] M. Smolik and V. Skala, "Vector field interpolation with radial basis functions," in *Proceedings of SIGRAD 2016, May 23rd and 24th, Visby, Sweden*, no. 127. Linköping University Electronic Press, 2016, pp. 15–21.

From subcritical behavior to a correlation-induced transition in rumor models

Guilherme Ferraz de Arruda,[1,*] Lucas G. S. Jeub,[2] Angélica S. Mata,[3] Francisco A. Rodrigues,[4] and Yamir Moreno[2,5,6]

[1] *CENTAI Institute, Turin 10138, Italy*

[2] *ISI Foundation, Via Chisola 5, 10126 Torino, Italy*

[3] *Departamento de Física, Universidade Federal de Lavras, 37200-900, Lavras, Minas Gerais, Brazil*

[4] *Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - Campus de São Carlos, Caixa Postal 668, 13560-970 São Carlos, SP, Brazil.*

[5] *Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza 50009, Spain*

[6] *Department of Theoretical Physics, University of Zaragoza, Zaragoza 50009, Spain*

Rumor and information spreading are natural processes that emerge from human-to-human interaction. Such processes have a growing impact on people's daily lives due to increasing and faster access to information, whether trusted or not. A popular mathematical model for spreading rumors, data, or news is the Maki–Thompson (MT) model [1]. In this model, individuals can be in one of three states: ignorant, spreader, or stifler. The spreading evolves through the contact between nodes defined by an undirected network. Our process is defined in continuous time as a collection of Poisson processes. If the contact is between a spreader and an ignorant, the second node will learn the rumor and become another spreader at rate λ. On the other hand, if the contact happens between a spreader and someone that already knows the rumor (spreader or stifler), then the spreader that initiated the contact will lose interest in the rumor, thus becoming a stifler at a rate α. Existing work based on first-order mean-field approximations suggested that this model does not have a phase transition with rumors always reaching a finite fraction of the population irrespective of the spreading rate [2–4]. Here, we show that a second-order phase transition is present in this model, which is not captured by first-order mean-field approximations. Since the MT model has infinitely many absorbing states, the critical point is the spreading parameter that separates the two scaling regimes. Before this point, the final number of stiflers when the process reaches an absorbing state does not scale with the system size, and hence its fraction goes to zero in the thermodynamic limit. After the critical point, the number of stiflers scale with the system size. This transition is shown in Fig. 1 (a) and (b), where we present the order parameter (the fraction of stiflers) and the time to reach the absorbing state, respectively. Moreover, we propose and explore a modified version of the Maki–Thompson model that includes a forgetting mechanism, where each stifler spontaneously becomes ignorant at a rate δ. This modification changes the Markov chain's nature from infinitely many absorbing states in the classical setup to a single absorbing state and allows us to use a plethora of analytic and numeric methods to characterize the model's behavior. In particular, we were able to provide an estimation of the critical point by accounting for the correlations between states. The accuracy of our approximation is shown in Fig. 1 (c). More importantly, we find a counter-intuitive behavior in the subcritical regime, where the lifespan of a rumor increases as the spreading rate drops, following a power-law relationship. These results are summarized in Fig. 1 (b) and (d). Specifically, in Fig. 1 (b), we present the time to reach the absorbing state for different sizes, demonstrating the power-law subcritical behavior. Complementary, in Fig. 1 (d), we compare a similar result with two analytical approximations, confirming the power-law subcritical behavior. This behavior implies that, even below the critical threshold, rumors can survive for a long time. Furthermore, using an asymptotic analysis where we scale the model's parameters, we were able to show that no phase transition is expected in the first-order mean field approximation. This result, together with our critical point estimations, emphasizes the role of correlations in the MT model phase transition and motivates further research on developing more sophisticated mean-field approximations. Together, our findings are at odds with most classical results and show that the dynamic behavior of rumor models is much richer than previously thought. Thus, we hope our results motivate further analytical and numerical research and investigations involving real-world systems. The work described in this abstract has been published in [5].

* gui.f.arruda@gmail.com

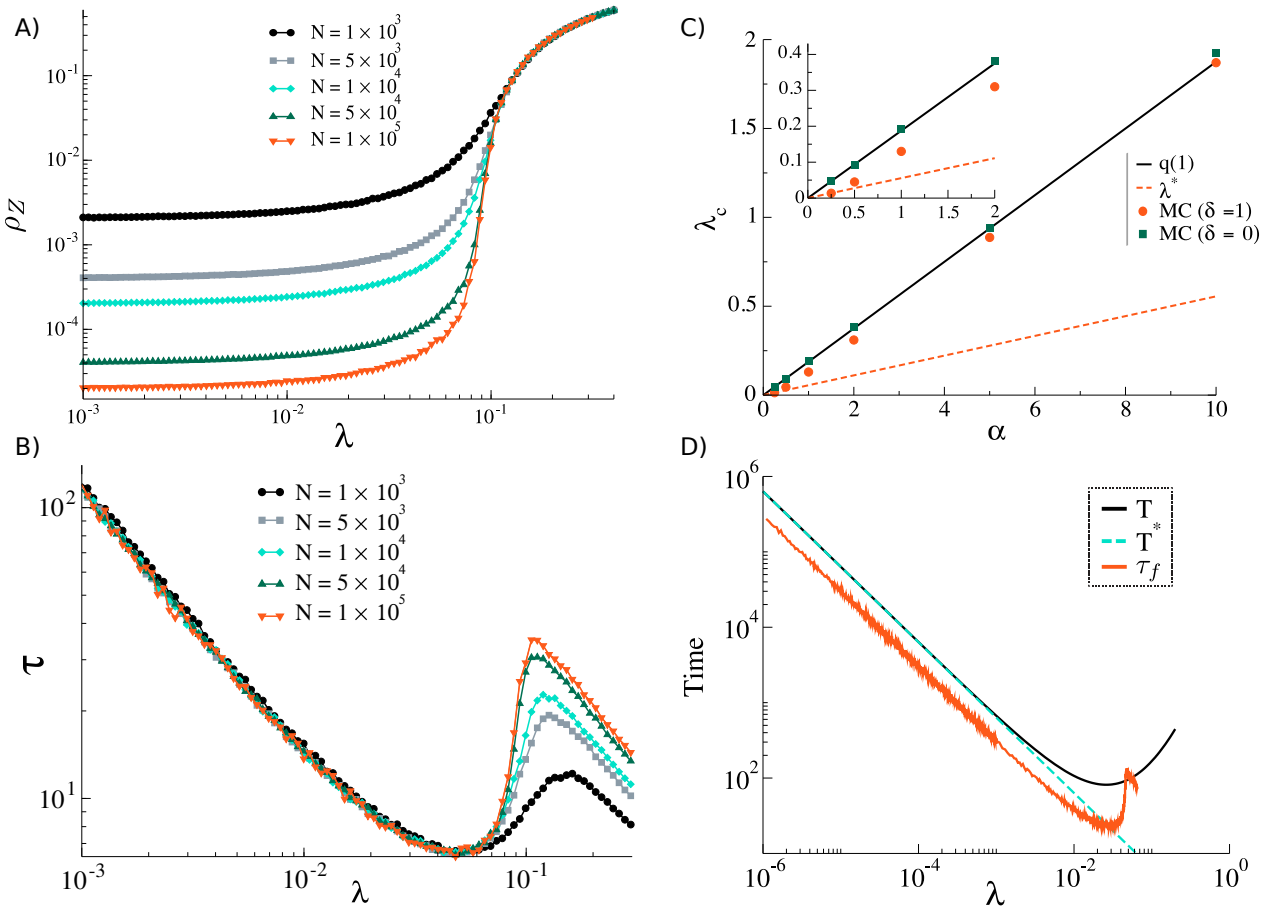


FIG. 1. **Summary of results showing the phase transition in the Maki–Thompson model.** In (a) and (b), we show the phase diagram and the time to reach the absorbing state for the standard MT model, $\alpha = 1$, and different sizes on a random regular network with $\langle k \rangle = 10$, respectively. In (c), we present the comparison between analytical and Monte Carlo critical point estimations for random regular networks with $\langle k \rangle = 10$ and $\delta = 1$ and $N = 10^6$. In the inset, we present the comparison for the low α regime. Finally, in (d), we show two approximations for the time to reach the absorbing state, T and T^*, as a function of λ for $\delta = 1.0$ and $\alpha = 0.5$. In this subfigure, the red curve results from Monte Carlo simulations in a random regular network with the same parameters, $\langle k \rangle = 10$ and $N = 10^5$.

[1] D. P. Maki and M. Thompson, *Mathematical models and applications* (Prentice-Hall Inc., Englewood Cliffs, N.J., 1973).

[2] Y. Moreno, M. Nekovee, and A. F. Pacheco, Phys. Rev. E **69**, 066130 (2004).

[3] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, Physica A: Statistical Mechanics and its Applications **374**, 457 (2007).

[4] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical processes on complex networks* (Cambridge University Press New York, NY, USA, 2008).

[5] G. Ferraz de Arruda, L. G. S. Jeub, A. S. Mata, F. A. Rodrigues, and Y. Moreno, Nature Communications **13**, 3049 (2022).

Global misinformation spillovers in the online vaccination debate before and during COVID-19

Jacopo Lenti[1], **Kyriaki Kalimeri**[1], **Andre Panisson**[2], **Daniela Paolotti**[1], **Michele Tizzani**[1], **Yelena Mejova**[1], **Michele Starnini**[1,3]
[1]ISI Foundation, Torino, Italy; [2]Centai, Torino, Italy
[3]Departament de Fisica, Universitat Politecnica de Catalunya, Barcelona, Spain
Keywords: *misinformation, social media, cross-national communication, public health*

Anti-vaccination views pervade online social media, fueling distrust in scientific expertise and increasing vaccine-hesitant individuals. Thus far, the scientific study of the debate around vaccination on online social media (OSM) has focused on specific countries [1, 2, 3, 4] or English-speaking users [5]. Upon arrival of COVID-19, it is now imperative to understand the flows of anti-vaccine — or *no-vax* — information not only nationally but internationally, in order to have a bird-eye view on the topic and inform effective communication campaigns. To address this need, in this work we build and analyze a series of international information flow networks by leveraging 316 million Twitter posts related to vaccines in 18 different languages from a pre-COVID era to April 2021. To this aim, we first investigate (i) how polarized, in terms of echo chambers phenomenon, the vaccination debate is in different countries, over time, to identify users in no-vax communities and (ii) how susceptible, in terms of circulation of information, are these no-vax communities to low quality information. We propose a flexible, language-neutral community detection approach, and combine it with human-in-the-loop expert knowledge to track polarization and echo chambers in different countries across time. Below, we highlight some of the results of this work.

We start by selecting 4 three-months periods before and during the pandemic, which we dubbed as i) pre-COVID, ii) pre-vaccine, iii) vaccine development, and iv) vaccine roll-out periods. We then select 28 countries in Europe, America and Oceania having at least 2000 unique users in each period. We construct the retweet (RT) networks corresponding to each country and time period, detect communities by using hierarchical clustering, and label a sample of tweets from each community to identify clusters of users exposed to no-vax content. We found 52 of these "no-vax" communities. Note that this does not imply all users in these communities hold anti-vaccination opinions, but that they are more likely to be exposed to such material. We find that no-vax communities are generally present in English-speaking countries, with respect to Spanish speaking ones. However, some of the relatively largest country-specific no-vax communities appear in France, Italy, Netherlands, Poland, and the United States.

Turning to potential echo-chambers in these networks, we first quantify the degree of polarization in the vaccination debate by using the Random Walk Controversy (RWC) score [6], which measures how much users in no-vax communities are exposed to information coming from their own side vs. the rest of the network. The RWC score is overall very high, indicating that **the vaccination debate is generally highly polarized**. However, it decreases substantially over time, suggesting that users in novax communities became less isolated in the vaccination discourse during the COVID pandemic. Secondly, we investigate whether the users in the no-vax communities are exposed to information sources different from the rest of users [7]. To this aim, we look at the content shared by the users, constructing a co-sharing (CO) network where users are connected by a link if they share the same URL, and gauge the similarity between the RT and CO networks by computing the Normalized Mutual Information (NMI) between their community structures. On average, the NMI of the networks with a novax community is higher than the others (0.27 vs 0.22, $p < 0.05$), indicating that users **in no-vax communities tend to have common information sources**. Some countries, such as the U.S. and Brazil, show an especially high NMI, indicating that the polarization in the retweet network is reflected in the different content shared.

Considering the behavior of **users in no-vax communities, we find that they are more likely to retweet, share URLs, and especially URLs to YouTube than other users**. Furthermore, the URLs they post are much more likely to be from low-credible domains (identified using lists of such domains in 4 languages), compared to those posted in the rest of the networks. The difference is remarkable: **26.0% of domains shared in no-vax communities come from lists of known low-credible domains, versus only 2.4% of those cited by other users** ($p < 0.001$).

Next, we investigate the effects of content moderation by Twitter on the vaccination debate. We find that the average proportion of suspended accounts in no-vax communities is much larger than the rest of users, for each country and period considered (average 13.3% vs 1.8%, $p < 0.001$). **A large portion of suspensions come after the January 2021 U.S. Capitol attack in Washington, D.C.**[1] These findings suggest that political leaning is often associated with strong stances taken in the vaccination debate (in line with previous literature [1, 4]) and that actions taken in the political domain may greatly impact the quality of the public health discourse.

Next, we quantify the information spillover across countries by considering the number of retweets from one country to another, normalized by the total number of retweets produced and received in the two countries (Fig 1a). We find

[1]The suspensions were announced by Twitter https://blog.twitter.com/en$_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--$

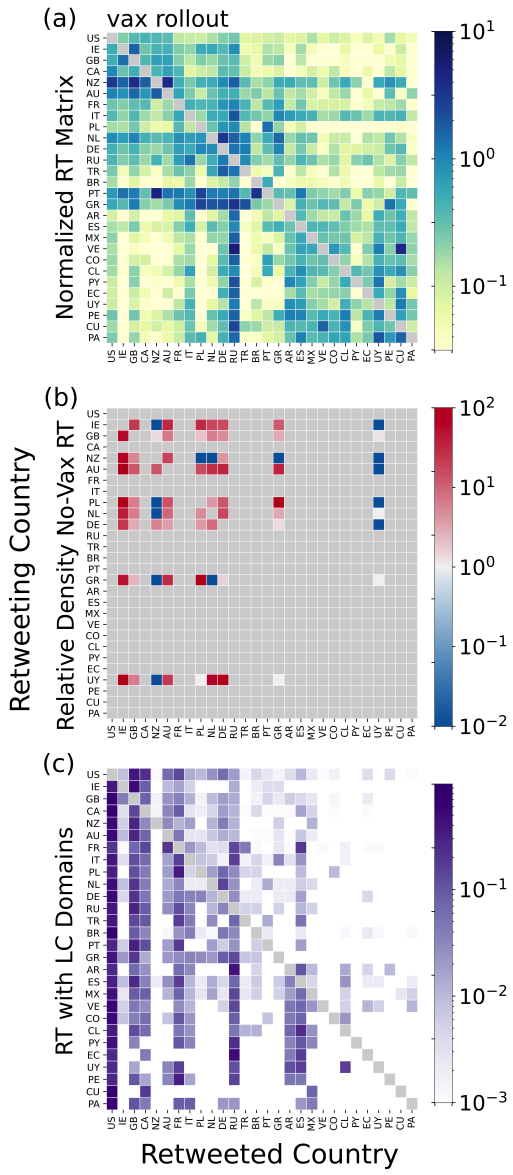


Fig. 1. **Cross-border information flows in the global vaccination debate for last period – vaccination rollout.** (a) Normalized number of retweets (excluding diagonal elements from the plot, colored in grey), (b) Probability of interaction between users in no-vax communities from one country to another, with respect to the interactions between other users from the same pair of countries (see Methods). Darker red (blue) elements of the matrices represent higher (lower) tendency of cross-border interactions between users in no-vax communities with respect to other users (countries without no-vax communities colored in grey), (c) Proportion of URLs that come from Retweeted Country among the low-credible domains imported by Retweeting Country (countries importing less than 10 low-credible URLs are coloured in grey). Element a_{ij} of each matrix represents information flow from country j to country i.

that the cross-border interaction matrices are not symmetric: information generally flows with a preferred direction. For instance, Spanish-speaking countries retweet English-speaking ones much more than the opposite. **The United States is central in the global information flow (despite flows being normalized)**, being a net exporter of information to the rest of the world. Interestingly, from pre-vax period, Russia is also a net exporter, especially to South American countries: some of the most used hashtags in pre-vaccine and vax development periods are #sputnikesesperanza and #sputnikparaelpueblo.

Next, we quantify the strength of cross-border interactions between users in no-vax communities with respect to the rest of users (Fig 1b). We find that cross-border interactions between users in no-vax communities are generally much stronger, sometimes by orders of magnitude, than interaction from the rest of users, creating a **tightly-knit global no-vax network**. In particular, users in no-vax communities of English-speaking countries, Germany, and the Netherlands are tightly connected in all periods. Conversely, users in no-vax communities from Cuba and Russia are quite isolated.

Finally, we focus on the misinformation flows across countries by considering the fractions of low-credible domains imported per country (Fig 1c). As in the previous case, the matrices show a clear asymmetry. **U.S. users act as global misinformation superspreaders to the rest of the world**: 68% of all low-credible URLs retweeted worldwide come from U.S. (average over the four periods), a proportion much higher than the total volume (42%) retweeted from U.S.. Interestingly, the fraction of low-credible URLs coming from U.S. dropped from 74% in the vax development period to 55% in the vax rollout. This large decrease can be directly ascribed to Twitter's moderation policy: 46% of cross-border retweets of U.S. users linking to low-credible websites in the vax development period came from accounts that have been suspended following the U.S. Capitol attack. Finally, despite not having a list of low-credible domains in Russian, **Russia is central in exporting misinformation in the vax rollout period, especially to Latin American countries**. In these countries, the proportion of low-credible URLs coming from Russia increased from 1% in vax development to 18% in vax rollout periods.

In conclusion, despite the platform's tweet flagging and removal policies around COVID-19, it is the bout of account suspensions around the Washington riots that made the most impact on the national and international spread of vaccine-related misinformation, suggesting that the political concerns elicit much stronger curbing of the freedom of speech than the health one. As interaction with vaccine hesitant social media content has been related to an increased delay of vaccination [8], the lack of action in the first three periods of study may have contributed to the unnecessary deaths of unvaccinated individuals (estimated to be in hundreds of thousands in the U.S. alone [9]). In ongoing work, we are developing quantitative tools in order to gauge to what extent anti-vaccination sentiments expressed on Twitter relate to politicization of the topic by the political actors within each country.

[1] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140, 2020.

[2] Giuseppe Crupi, Yelena Mejova, Michele Tizzani, Daniela Paolotti, and André Panisson. Echoes through time: Evolution of the italian covid-19 vaccination debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 102–113, 2022.

[3] Mauro Faccin, Floriana Gargiulo, Laëtitia Atlani-Duault, and Jeremy K Ward. Assessing the influence of french vaccine critics during the two first years of the covid-19 pandemic. *arXiv preprint arXiv:2202.10952*, 2022.

[4] Matt Motta, Dominik Stecula, and Christina Farhart. How right-leaning media coverage of covid-19 facilitated the spread of misinformation in the early stages of the pandemic in the us. *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2):335–342, 2020.

[5] Ana Lucia Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on Facebook. *Vaccine*, 36(25):3606–3612, 2018.

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying Controversy in Social Media. In *WSDM '16: 9th ACM International Conference on Web Search and Data Mining*, pages 33–42, 2016.

[7] Bjarke Mønsted and Sune Lehmann. Characterizing polarization in online vaccine discourse—a large-scale study. *PloS one*, 17(2):e0263746, 2022.

[8] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[9] Krutika Amin, Jared Ortaliza, Cynthia Cox, Joshua Michaud, and Jennifer Kates. Covid-19 mortality preventable by vaccines. *Health System Tracker*, 2022.

Hybridization of Chemoinformatic and Bioinformatic Information Networks for Graph Diffusion of Drug-Target Interactions

Bruno Kaufman[1] and Ariel Chernomoretz[1,2]

[1]Fundación Instituto Leloir
[2]Physics Department, FCEN, University of Buenos Aires/IFIBA (CONICET)

Abstract

Drug development is a costly, time-consuming process. It is therefore convenient to repurpose existing drugs for novel treatments. Massive amounts of information exist on the topic of drug-protein interactions. In order to extract new insights from this wealth of data, massive computational approaches are warranted, combining bioinformatics and chemoinformatics. In our work, we explore the combination of several sources of data: drug-protein interactions and chemical similarity measures based on molecular fingerprints (Tanimoto similarity) and drug scaffold hierarchies. These come in the form of complex networks, and a hybrid network is obtained by combining them. In this contribution we present a network diffusion strategy to predict new drug-protein interactions.

Keywords — network diffusion, drug repurposing, structural similarity

1 Introduction

Drug repurposing is a problem which seeks to use prior knowledge to infer new possible uses for existing drugs. One way of going about this task is to use data on drug-protein interactions, and attempt to predict new interactions of the same variety. Proteins are, in this context, referred to as "drug targets".

2 Methods

In our work, we make use of TDR-targets, a curated drug-target interaction database (Urán Landaburu et al., 2020). Our goal is to recover excluded parts of this database by combining the remaining information with chemical similarity data between drugs and making use of network diffusion.

Our procedure uses three separate graphs, each constructed with its own measure of drug-drug similarity: (1) Drug similarity by shared targets; (2) Chemical similarity by Tanimoto score; (3) Sharing of chemical scaffolds by Bemis-Murcko criterion (Bemis and Murcko, 1996).

These three networks are combined using a weighted linear combination of their edge values. Each of these contributes to our hybrid, combined network through a separate weight parameter ($\vec{w} = (w_{targets}, w_{Tanimoto}, w_{scaffold})$), which should be learned through training. For a given hyperparameter set $\vec{w}$, we considered the unweighted Laplacian of the integrated network to perform a label propagation procedure (Chapelle et al., 2009).

3 Results

Our integrated network is composed by 5k target and 7M drug nodes linked by 100M drug-drug and 2M drug-target edges. In order to implement a performant prioritization methodology we partitioned the chemical space into Louvain drug communities and looked for cluster specific hyperparameters $\vec{w}$) during the training phase. The rationale for this methodology was that the relative importance of each knowledge layer could be differentially adjusted, better

reflecting the interplay between the propagation of information from known drugs and network topology for different parts of the network.

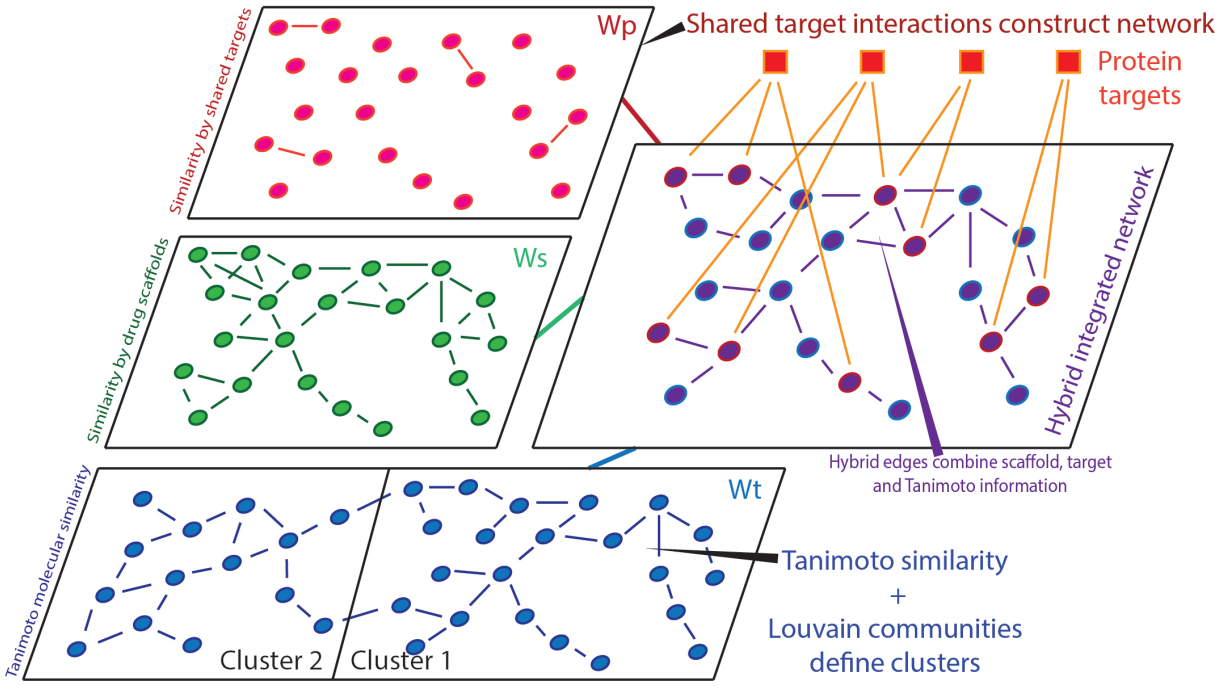


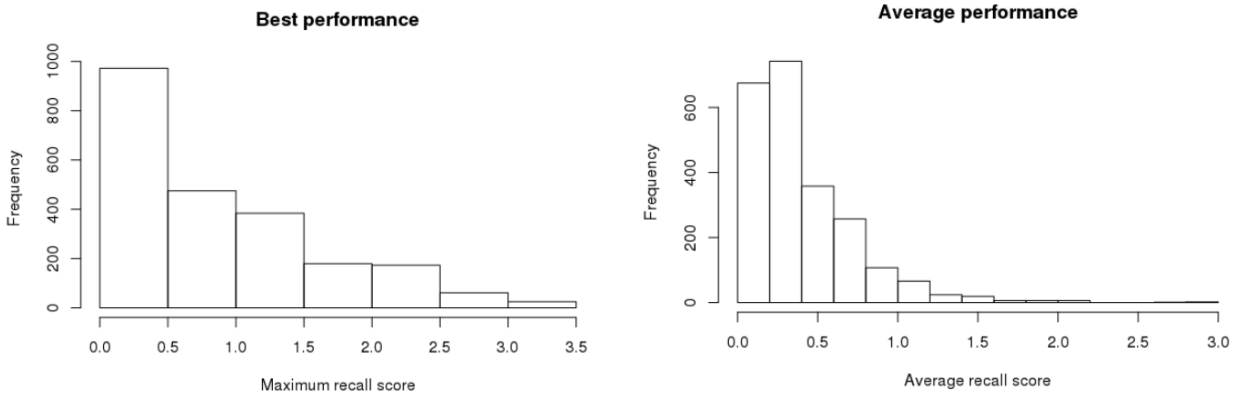
Figure 1: Architecture of information used. Drug targets are depicted as orange squares. According to how many targets a given drug shares, they may be linked by target similarity (top left, red). The same drugs may also be linked if they are comprised of similar scaffolds (middle left, green). Tanimoto similarity also links drugs to each other (bottom, blue); this measure is also used to define Louvain communities. In the example depicted, network diffusion is applied on Cluster 1, using active drugs as seeds (circles with orange perimeter). The parameter balance (W_p, W_s, W_t) is varied until an optimal combination is found in terms of recovery capability. After this, each target is prioritized individually, and their performance is assessed.

Using these locally optimized models we were able to evaluate the prioritization performance for targets on the validation set, according to their recovery capability in terms of the number of true positive elements found within the top 5,10,20,50 ranked elements of the prioritization list. Our score is determined as follows:

$$S_{t,c} = \sum_{n}^{(5,10,20,50)} \frac{E_n}{n}$$

where E_n is the number of elements of the validation set that were found within the top n elements of the ranking.

Table 1: Distribution for the performance of individual targets. This score reflects the recovery of excluded interacting drugs within the top 5, 10, 20 and 50 recommendations.



A target may have different performance in different clusters. Thus we analyze two measures of performance: maximum score among all clusters, and average score among them. Out of 2.2k targets, only around 200 have a performance of zero. The remaining ones are shown in Table 1. To this date, this represents a coverage of roughly 40% of our 5k drug targets. Results are promising, showing efficient use of available information that results in a 5-fold increase in reliable predictions.

References

L. Urán Landaburu, A. J. Berenstein, S. Videla, P. Maru, D. Shanmugam, A. Chernomoretz, F. Agüero, Tdr targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration, Nucleic acids research 48 (2020) D992–D1005.

G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. molecular frameworks, Journal of Medicinal Chemistry 39 (1996) 2887–2893. URL: `https://doi.org/10.1021/jm9602928`. doi:10.1021/jm9602928. `arXiv:https://doi.org/10.1021/jm9602928`, pMID: 8709122.

O. Chapelle, B. Scholkopf, A. Zien, Eds., Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews], IEEE Transactions on Neural Networks 20 (2009) 542–542. doi:10.1109/TNN.2009.2015974.

Identification and Mapping of Vehicle Robbery and Theft in Rio de Janeiro

Douglas Ferreira[1,2], Jennifer Ribeiro[2,3], Renato Freitas[2], André Pimenta[2], Valter Felix[2], and José Mendes[1]

[1] *Department of Physics, University of Aveiro, Aveiro, Portugal*
[2] *LISComp – Laboratory, Federal Institute of Rio de Janeiro, Paracambi, RJ, Brazil*
[3] *Department of Geophysics, National Observatory, Rio de Janeiro, RJ, Brazil*

Keywords: Complex networks; Crime; Rio de Janeiro, Nonextensive statistical mechanics

Among the various problems in human society, crime is one of the most worrying and harmful since it strongly interferes with the sustainable development of society. In a highly tragic way, Rio de Janeiro (Brazil) has alarming rates of crime and violence, of all possible types, from petty offenses to the most tragic cases of robberies and murders.

Empirical evidence shows that crime has a remarkable regularity of concentration in several dimensions that relate to the context (target, location, offender, etc.) and characteristics (spatial, temporal, type of crime, etc.) [1]. Thus, in the present work, the characterization of activities of robberies and theft of vehicles that occurred in the city of Rio de Janeiro was carried out to develop methods and models to evaluate spatial concentrations of crimes that collaborate in the search for existing patterns among the criminal activities that occurred.

Given the successful application of statistical physics methods in modeling and describing various social systems (such as interactions between individuals [2], epidemic propagation [3], and human cooperation [4]), in this work, we analyzed the behavior of crime from the point of the theory of complex networks, analyzing the spatial and temporal dynamics of crimes. It is worth mentioning that in complex systems, individual interactions between the system elements produce global behaviors that cannot be inferred only by individual analyses. It happens since the relationships between the system components create non-linear complex behaviors, which in many cases translate into long-range spatial and temporal interactions. Such methods have already proven effective in crime studies for certain political or geographical contexts [6,7]. In this way, we use complex network techniques for mining and analyzing criminal data, helping to understand important issues related to the "criminal phenomenon" and contributing to a clearer and broader view of the behavior of criminal activities and how they are related and/or evolve.

The data used in the present work have information on the neighborhood, the date, and the time when each crime occurred. Those data were provided by the Rio de Janeiro Public Security Institute (ISP-RJ). The types of crimes analyzed were classified as "vehicle theft" (VT) and "vehicle robbery" (VR). The data provided have information from 2010 to 2015, with 42,365 vehicle thefts and 69,330 vehicle robberies.

For the construction of crime networks, we will consider that each neighborhood in Rio de Janeiro will become a vertex of the network if it has at least one occurrence of the considered type of crime in the range from 2010 to 2015. The connections between the vertices will be carried out according to the following method. A time window of size W is defined and inserted in the chronologically organized data to connect all the vertices inside this window. The time window is inserted starting from the first vertex, causing this vertex to be connected to all vertices that are inside this window. After that, the window is moved forward, starting at the next occurrence. This second occurrence is then connected to all subsequent events that fall within the time window, after which the window is again moved forward to the next occurrence of the considered crime. This binding process is repeated until moving the window forward is no longer possible. This network construction model works as a temporal filter for the connections between vertices, allowing vertices that are close in time to be connected, regardless of whether they are immediately

subsequent or not. Likewise, this model prevents the connection of vertices that are very distant in time and thus have a low probability of being correlated.

A fundamental step in creating the time window network is to define an "ideal value" for the window size, W, since, *a priori*, there is no preferred or pre-defined value. Thus, it is necessary to develop a methodology capable of finding the best value for W. To obtain this value, we use the concept of *community* and apply it directly to the data time series. We vary the time window value for each type of crime and observe that the number of communities reaches a maximum value for a specific time window value (Fig.1). This specific time window value was the one considered for constructing the networks for VT and VR occurrences.

Considering the network as *weighted*, where the weight of each link is the number of times a given link is repeated, in Fig. 2 we have the distribution of strengths for the VT and VR networks. The results present the characteristic of the behavior described by the non-extensive statistical mechanics [8], characterized by the emergence of probability distributions in q-exponential functions,

$$P(\geq s) = -e_q(\beta s) = [1-\beta(1-q)s]^{1/(1-q)}, \tag{1}$$

where s is the strength value of each vertex and β is the positive constant for each network.

Fig.3 shows how are distributed the connections between the vertices of the networks for both VT and VR crimes. In the network maps, the vertices have the geographic locations of each neighborhood, and only the edges with a strength greater than 50 are represented. It means that the edges shown on the map only refer to crimes that occurred at least 50 times between one neighborhood and another and within a time interval equal to or less than the time window, W.

The networks were also organized to allow groupings of neighborhoods by their respective regions. It was done to provide another perspective for visualizing the connections between neighborhoods. The results are shown in Fig.4, where only the 40 neighborhoods with the highest number of occurrences are highlighted.

Our results indicate that theft and robbery of vehicles have strong spatiotemporal correlations, meaning that each occurrence cannot be evaluated and analyzed as an isolated event. Furthermore, our study gives a first step to correlating criminal activities in different areas of Rio de Janeiro.

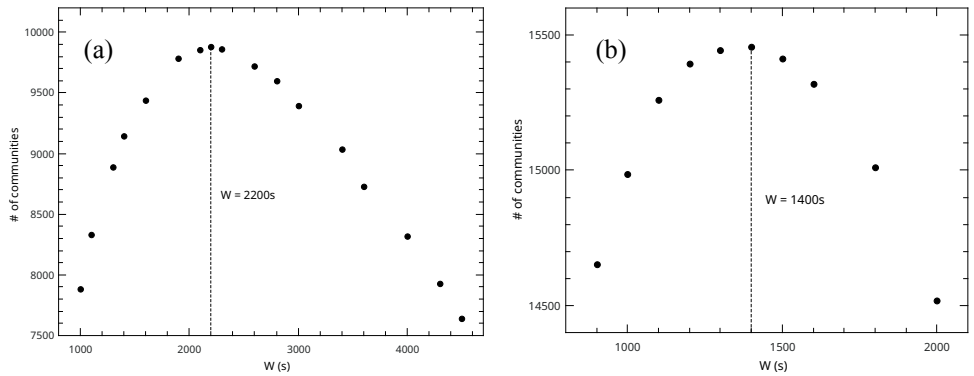


Fig.1- Number of communities for each time window value, W, for vehicle (a) theft and (b) robbery.

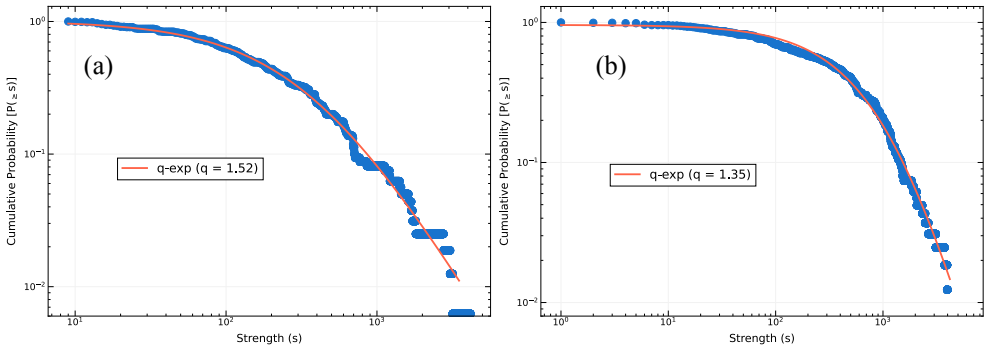


Fig.2- Distributions of strengths for vehicle (a) theft and (b) robbery networks. The solid lines correspond to adjustments according to equation (1) of the text.

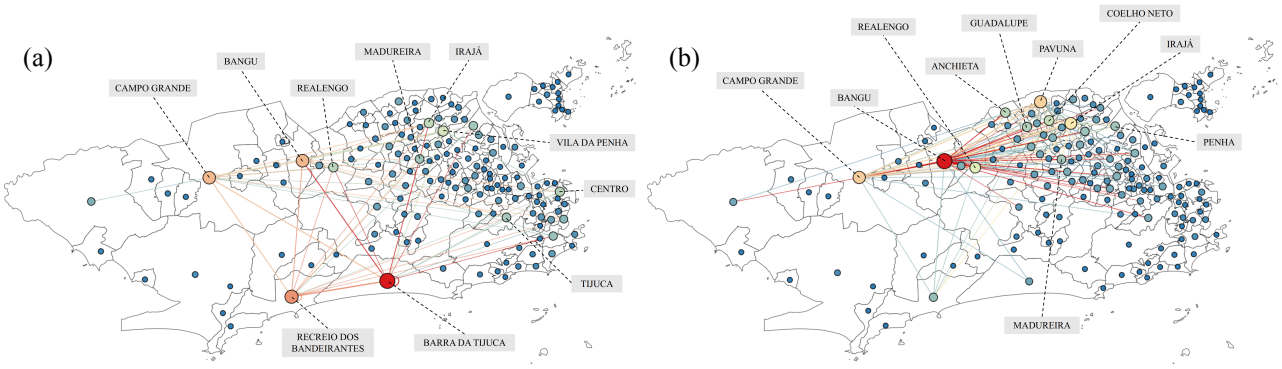


Fig.3- Network map for vehicle (a) theft and (b) robbery, from 2010 to 2015, where each edge has $s \geq 50$. The figure highlights the 10 neighborhoods with the highest *strength* value.

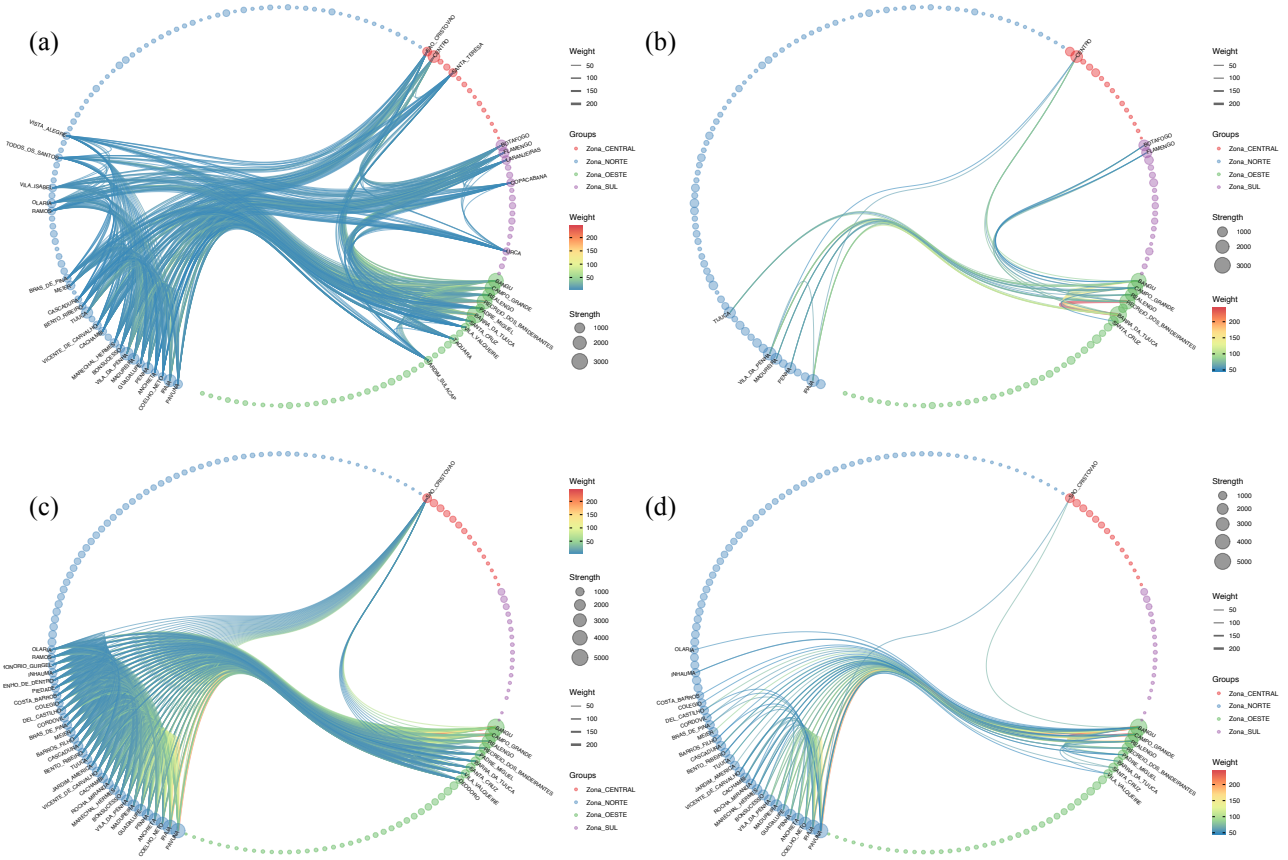


Fig.4- Networks divided by zones. The minimum weight values and vehicle type of crimes considered were: (a) 0, theft; (b) 50, theft; (c) 0, robbery; (d) 50, robbery.

References

[1] G. Farrell. *Crime Prevention & Community Safety*, vol.17, n.4, (2015) 233–248.
[2] C. Castellano, S. Fortunato, V. Loreto, Rev. *Modern Phys*. 81 (2009) 591.
[3] R. Pastor-Satorras *et al.*, *Rev. Modern Phys*. 87 (2015) 925.
[4] M. Perc *et al., Phys. Rep*. 687 (2017) 1–51.
[5] G. Spadon *et al.*, *Inf. Technol. Gener. - Springer*, (2018) 493–500.
[7] F. Calderoni, D. Brunetto, C. Piccardi, *Soc. Networks.* 48 (2017) 116–125.
[8] D. A. Bright *et al., J. Contemp. Crim. Justice*. 31 (2015) 262–278.
[9] C. Tsallis, *Journal of statistical physics.* 52.1 (1988) 479-487.

Identifying vaccine-mechanism bias in mathematical models of vaccine impact: the case of tuberculosis

Mario Tovar[1,2], Yamir Moreno[1,2,3] & Joaquín Sanz[1,2]

1. Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza 50009, Spain.
2. Department of Theoretical Physics, University of Zaragoza, Zaragoza 50009, Spain.
3. Centai Institute S.p.A, Corso Inghilterra 3, 10138 Torino, Italy

Keywords: Epidemiology, Tuberculosis, Vaccines, Bayesian Inference.

Background: Despite the decay in Tuberculosis (TB) incidence and mortality achieved worldwide since 1990 [3], its yearly rate of reduction is arguably too slow to meet the goal settled by the World Health Organization (WHO) in the End-TB strategy, which consists of completing a reduction of TB incidence and mortality rates by 90% and 95%, between 2015 and 2035 [8]. Instead, during 2020, and for the first time in decades, the world suffered an increase in global TB burden levels with respect to previous years, while, during that same year, the WHO estimated that TB was the cause of death of more than 1.5 million people worldwide, combining HIV negative and positive cases [5].

The cause of this increase was the irruption of the COVID-19 pandemics, which threatens, in countries like India or Indonesia, to raise the TB death toll back to even higher levels in the next few years [1, 7]. This issue, alongside with the ever-increasing rates of emergence of drug resistance [4], evidence the need of new epidemiological interventions and tools against TB. Among those tools and interventions, the development of a new and better vaccine than the current bacillus Calmette-Guerin (BCG) seems necessary, as in the former, the efficacy against the more transmissible respiratory forms of the disease in young adults is disputed [2].

For such a task, specially on TB, where several candidates are under development, a robust impact forecast which is based on epidemiological model arises as a powerfully tool to help evaluating those vaccines before introducing them in the general population. Nonetheless, in the development of vaccines against TB, a number of factors represent burdensome difficulties for the design and interpretation of randomized control trials (RCTs) of vaccine efficacy. Among them, the complexity of the transmission chain of TB allows the co-existence of several routes to disease that can be observed within the populations from where vaccine efficacy trial participants are sampled. Ultimately, this makes it difficult to derive mechanistic descriptions of the vaccines in terms of the mathematical model if only trial-derived readouts of vaccine efficacy are used. This happens since, intuitively, the same efficacy readouts may lean on the ability of a vaccine to arrest only some, but not all, the possible routes to disease. This increases uncertainty in evaluations of vaccine impact based on transmission models, since different vaccine descriptions of the same efficacy readout typically lead to different impact forecasts.

Methods: Aiming to address some of the difficulties in translating real RCTs results to spreading models, in this work, we develop a Bayesian framework to evaluate the relative compatibility of different vaccine descriptions with the observations emanating from a randomized clinical trial of vaccine efficacy. This offers an unbiased framework to estimate vaccine impact even when the specific mechanisms of action of the given vaccine are not explicitly known, providing a more agnostic impact evaluation of those vaccines.

The method we propose combines *in-silico* trials of the real RCT with a Bayesian framework that allows to capture the realtive compatibility of the vaccine descriptions with the real outcome of the trial. For this, we first proposed 7 different vaccine descriptions that capture the three routes to disease that might be observed in TB and in a IGRA-positive trial, which are related to natural TB history, and then simulated each one of them in the context of the RCT. Then, we compute the relative compatibility of each vaccine descriptions with the real results along with

the most probable intrinsic-efficacy value (ε) which can be used directly in the spreading models to model such a vaccine. Using our methodology, we analyzed the results reported for the vaccine M72/AS01$_E$ clinical trial as a case study [6].

Results: We applied our bayesian framework to the case study in order to weight the relative compatibility of each one of the possible mechanistic effects of the vaccine that might be observed in an IGRA-positive trial. Figure 1 shows the core results of this procedure in which, first, we produced *in-silico* simulations for a myriad of possible intrinsic-vaccine efficacy (those are the spreading model-related efficacies of the vaccine) in the RCT, which are reported in the clouds. Then, we applied the Bayes rule to those clouds for deriving the posterior probabilities of both the intrinsic efficacy values, that are introduced later in spreading models, and the relative compatibility of the whole mechanistic decription with the real efficacy of $\text{VE}_{dis} = 49.7\%$ observed in the trial.

Those posterior probabilities, that are shown in Figure 1 B and C, enables the possibility to forecast the impact of the vaccine using an spreading model, as we get the most compatible value of ε, that captures the intrinsic efficacy of the vaccine in the model, along with the overall compatibility of this description with the real outcome of the trial. Moreover, we derived an agnostic estimate of the impact of the vaccine using the posterior probabilities reported in Figure 1 B to produce a weighted forecast that is agnostic to the mechanistic description of the vaccine.

Conclusions: This work enlightens the problem of translating the outcome of TB vaccine's RCTs to the spreading models while aiming to produce robust forecast impact evaluations. We shown here that, even in cases with high uncertainty, such as the case study, a clean-well designed procedure allows to disentangle the effect of the vaccine at the mechanistic level. This, ultimately, leads to the possibility of producing agnostic impact evaluation of new vaccines, which reduces the gap betweeen reality and models. Moreover, the type of RCTs considered here, conducted on IGRA-positive individuals, emerged as a promising design architecture after the encouraging results reported for the vaccine M72/AS01$_E$ clinical trial, as they might also be analyzed with this -or a similar- method to improve our knowledge of the effect of a new vaccine, both at the biological level, at the model level, and at the population level.

References

[1] Lucia Cilloni et al. "The potential impact of the COVID-19 pandemic on the tuberculosis epidemic a modelling analysis". In: *EClinicalMedicine* 28 (2020), p. 100603.

[2] Paul EM Fine. "BCG: the challenge continues". In: *Scandinavian journal of infectious diseases* 33.1 (2001), pp. 58–60.

[3] Anoushiravan Kazemnejad et al. "Global epidemic trend of tuberculosis during 1990-2010: using segmented regression model". In: *Journal of research in health sciences* 14.2 (2014), pp. 115–121.

[4] Christoph Lange et al. "Drug-resistant tuberculosis: an update on disease burden, diagnosis and treatment". In: *Respirology* 23.7 (2018), pp. 656–673.

[5] World Health Organization et al. *Global tuberculosis report 2020. Geneva: WHO; 2020.* 2021.

[6] Dereck R Tait et al. "Final analysis of a trial of M72/AS01E vaccine to prevent tuberculosis". In: *New England Journal of Medicine* 381.25 (2019), pp. 2429–2439.

[7] Mario Tovar et al. "Modeling the impact of COVID-19 on future tuberculosis burden". In: *Communications medicine* 2.1 (2022), pp. 1–10.

[8] Mukund Uplekar et al. "WHO's new end TB strategy". In: *The Lancet* 385.9979 (2015), pp. 1799–1801.

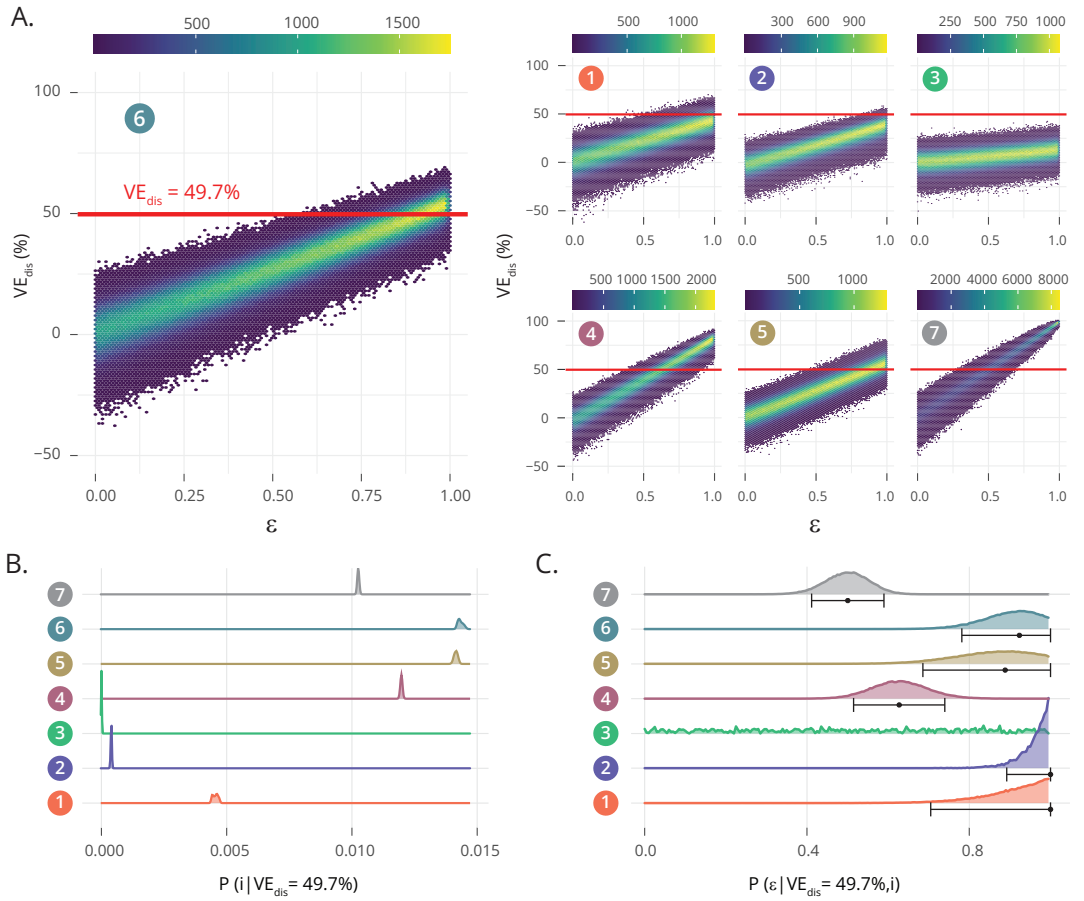


Figure 1: Bayesian analysis of possible modeling architectures underlying a trial-derived observation of vaccine efficacy. **A:** Absolute frequency density clouds of efficacy values VE_{dis} obtained in sets of $N = 2 \cdot 10^6$ clinical trial simulations per model, uniformly distributed across the intrinsic vaccine efficacy parameter ε(10000 points for each ε value, yielding $N = 2 \cdot 10^6$ points inside the cloud). Red horizontal lines mark the PoD efficacy observed in the M72/AS01$_E$ trial $\mathrm{VE}_{dis} = 49.7\%$. **B:** Marginal posteriors $P(i|\mathrm{VE}_{dis} = 49.7\%)$, capturing the relative compatibility of each model with respect to the efficacy observed in the M72AS01E trial. **C:** Distribution $P(\varepsilon|\mathrm{VE}_{dis} = 49.7\%, i)$ of the intrinsic vaccine efficacy parameter ε in each model type, given the observed efficacy $\mathrm{VE}_{dis} = 49.7\%$, along with mean and 95% confidence intervals associated to them. For M3, the CI was omitted, for it spans the entire range $\varepsilon \in [0, 1]$, as the model fails systematically to produce simulation instances compatible with the observed $\mathrm{VE}_{dis} = 49.7\%$.

Impact of COVID-19 on Chile's Internal Migration

Erick Elejalde
L3S Research Center, Germany

Victor Navarro
Data Science Institute, UDD, Chile

Loreto Bravo
Data Science Institute, UDD, Chile

Leo Ferres
Data Science Institute, UDD + Telefónica R&D, Chile
ISI Foundation, Italy

Keywords: internal migration, mobile data, COVID-19 pandemic, human mobility

Abstract

So far, most research regarding mobility and COVID-19 has focused either on studying day-to-day mobility to understand and measure the efficacy of lockdowns, on correlating mobility and socioeconomic factors bearing on new cases and deaths, or on building origin-destination matrices to inform epidemiological models [2]. The present report addresses a less-studied phenomenon of general mobility, namely, the patterns of permanent (or at least long-term) relocation within a country during the pandemic. The phenomenon of permanent relocation can be studied as a special kind of voluntary migration, where people leave their homes looking to improve their quality of life rather than forced migration (e.g., fleeing famine or political persecution) [1].

We focus on the Metropolitan Region (MR) of Santiago, Chile. Despite occupying a relatively small area, this region hosts a population of over eight million inhabitants, considering the projections of the 2017 census[1]. This makes it comparable in sheer numbers to other major metropolitan areas such as Hong Kong, Baghdad, or New York City. The MR is administratively divided into six provinces and 52 comunas[2]. We approach the study of this voluntary relocation phenomenon by quantifying how the population of each comuna of Santiago migrated out of the city and how they were distributed to the other regions of Chile. To understand the potential impact of the pandemic, we compare the migration patterns against the year 2017.

For each year (i.e., 2017 and 2020), we analyzed eight months (March 1 until November 30) of eXtended Detail Records (or XDRs) for approximately 1.3M devices in Santiago. These XDRs can be thought of as a tuple $< n, A, d >$, where n is an anonymized (hashed) mobile phone number that connects to a cell tower at latitude/longitude A on the day and time d. Each device was assigned a home antenna in one of the administrative areas (i.e., comunas) using those records between 7 pm and 7 am during weekdays [3]. We use the week of March 9 until March 15, 2020, taken as the business-as-usual week before the implementation of lockdowns (March 16) and full quarantines (May 16). Since classes had started the week before, we assume that most people would be at their primary

[1] `http://www.censo2017.cl/`
[2] `https://en.wikipedia.org/wiki/Santiago_Metropolitan_Region`

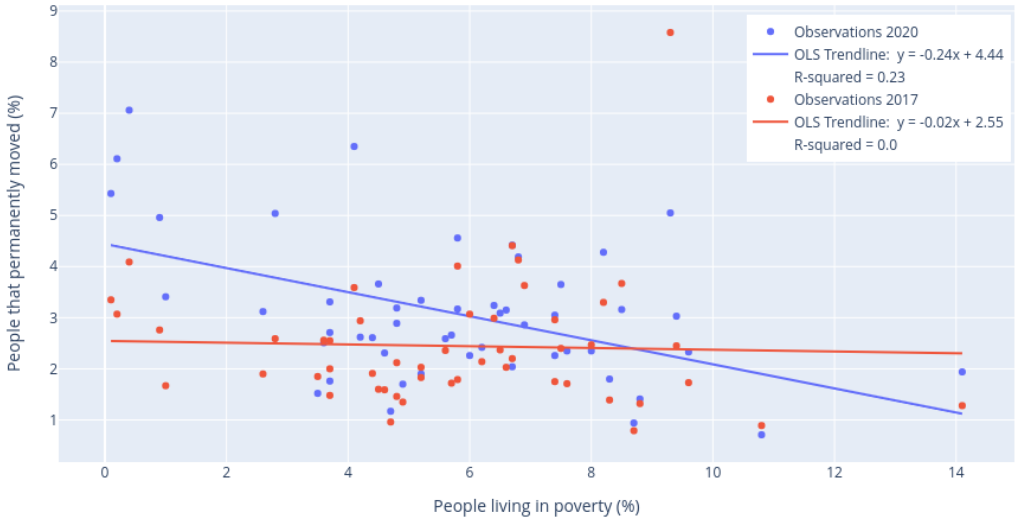


Figure 1: Emigration per comuna from the Metropolitan Region in 2017 (red) and 2020 (blue).

residence. We operationalized permanent migration as follows: for each device, we calculated home location per week after the baseline (i.e., 03/16/2020) until November 30, 2020, and if the statistical model of home in the four weeks of November was outside the Metropolitan Region, we assumed that the device had moved permanently. Correspondingly, we apply the same methodology to 2017. To validate our operationalization of migration, we compare our measurements from 2017 against the corresponding information from the national census conducted in Chile that same year. We found that our model of migration based on mobile data can be used to approximate the distribution of movement from the MR to other regions in Chile according to the census results ($r(13) = .93, p < .001$).

Our preliminary results show that about 173.8K people (2.17% of the projected 8.1M population) permanently left the MR in 2020. In contrast, an estimated 110.7K moved in the same period of 2017. However, the preferences of destinations stayed relatively stable. The main difference was in the Valparaíso region (a neighboring region and a popular vacation destination), which saw a significant increase in immigration from the MR. Another relevant difference during the pandemic compared to 2017 is the comunas of origin for the people who moved. The average socioeconomic level appears to be critical in explaining migration patterns during the pandemic. While for the 2017 dataset, this feature does not show any correlation with the percentage of the population migrating from each comuna, in 2020, the "percentage of people living in poverty"[3] alone explains 23% of the variance for the modeled variable (see Figure 1).

Regarding the destination, some regions saw significant differences in the distribution of the origin of the immigration. Figure 2 shows the normalized percentage difference by destination between 2017 and 2020. For example, of the total number of people who migrated to Los Lagos in 2020, 11.34% came from Las Condes (one of the wealthiest comunas in the MR), compared to 3.71% in 2017, for an increase of 7.63% (z-score=4.70). In contrast, for the same region in 2020, only 2.23% came from San José de Maipo (a middle-class comuna), compared to 4.95% in 2017, for a decrease of 2.72% (z-score=-2.15). Once again, the results suggest that the socioeconomic status of the comunas of origin played an essential role in the movement during the pandemic. Richer comunas like Las Condes and Providencia significantly increased their contribution to multiple regions.

[3] `www.comunidadescolar.cl/wp-content/uploads/2019/10/I%CC%81NDICE-DE-POBREZA-POR-COMUNA-2017.pdf`

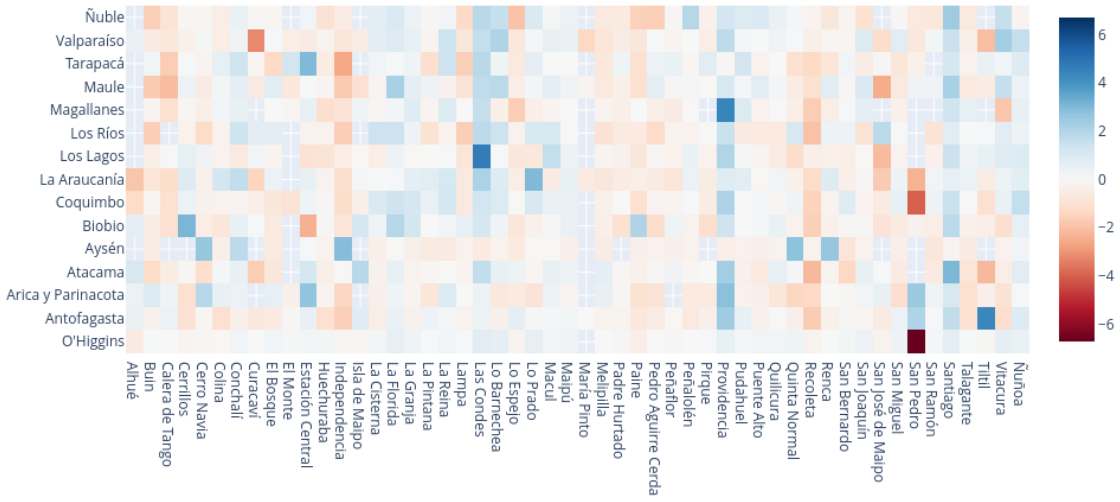


Figure 2: Percentage difference on the destination regions per origin comuna 2017 - 2020 (z-score normalization).

For 2020, We also calculate the difference between origin and destination for a combined index that represents the relative quality of life in urban areas (ICVU[4]). Although some comunas with a better quality of life necessarily had to cede when moving, we found that not everyone was willing to sacrifice in the same way. For example, Providencia (second in the ranking) conceded more than expected, while Vitacura (first in the ranking) was considerably more conservative, staying well below the trend. A similar tendency can be seen in the difference in the percentage of poverty between the origin and destination. People tended to move to communes with similar levels of poverty. Regarding migration from urban to rural areas, people predominantly stayed in urban comunas. Only poorer comunas showed a small percentage ($< 10\%$) of the migrants moving to rural areas.

Although the decision to migrate is complex and multifaceted, our findings suggest that during the COVID-19 pandemic, an advantageous socioeconomic status might have facilitated relocation. This gap in access to migration during periods of crises based on socioeconomic factors is a serious problem and requires further study to understand its underlying mechanisms better.

References

[1] Elizabeth Colson. Forced Migration and the Anthropological Response. *Journal of Refugee Studies*, 16(1):1–18, 03 2003.

[2] Nicolò Gozzi, Michele Tizzoni, Matteo Chinazzi, Leo Ferres, Alessandro Vespignani, and Nicola Perra. Estimating the effect of social inequalities on the mitigation of covid-19 across communities in santiago de chile. *Nature communications*, 12(1):1–9, 2021.

[3] Luca Pappalardo, Leo Ferres, Manuel Sacasa, Ciro Cattuto, and Loreto Bravo. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ data science*, 10(1):29, 2021.

[4] `https://estudiosurbanos.uc.cl/10-anos-calidad-de-vida-urbana-icvu-2020/`

Influence maximization in Boolean networks

Thomas Parmer and Filippo Radicchi

Center for Complex Networks and Systems Research,

Luddy School of Informatics, Computing, and Engineering,

Indiana University, Bloomington, Indiana 47408, USA

Luis M. Rocha

Consortium for Social and Biomedical Complexity,

Systems Science and Industrial Engineering Department,

Thomas J. Watson College of Engineering and Applied Science,

Binghamton University (State University of New York), Binghamton, New York 13902, USA and

Instituto Gulbenkian de Ciência, Oeiras 2780-156, Portugal

The optimization problem aiming at the identification of minimal sets of nodes able to drive the dynamics of Boolean networks toward desired long-term behaviors is central for some applications, as for example the detection of key therapeutic targets to control pathways in models of biological signaling and regulatory networks. Unfortunately, the complexity of the optimization problem is exponential, making it exactly solvable on very small systems only. Some scalable approaches exist but they rely on linear approximations; other approaches estimate nonlinear effects but they are generally not scalable. In this talk, we introduce an alternative method inspired by those used in the solution of the well-studied problem of influence maximization for spreading processes in social networks. The computational time of the proposed method scales cubically with the network size. This is achieved thanks to some strong approximations, as for example neglecting dynamical correlations among Boolean variables. However, the method has the desirable feature of fully accounting for the nonlinear nature of Boolean dynamics. We validate the method on small gene regulatory networks whose dynamical landscapes are known by means of brute-force analysis (Figure 1). We then systematically apply it to a large collection of gene regulatory networks revealing that for about 65% of the analyzed networks, the minimal driver sets contain less than 20% of their nodes.

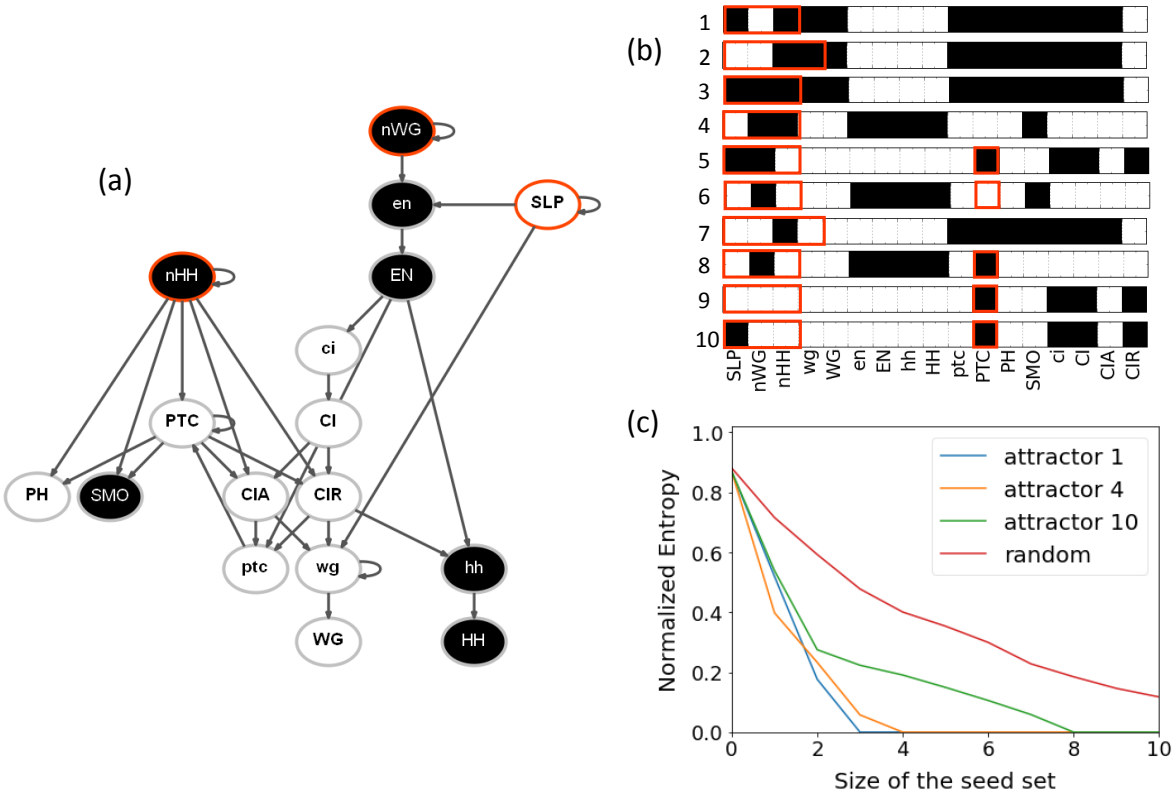


Figure 1: Driving Boolean networks to the desired attractor. (a) We consider the *Drosophila melanogaster* segment polarity network. In the visualization, a directed connection from node i to node j indicates that the state of node i is one argument of the Boolean function F_j that regulates the dynamical evolution of node j. In the visualization, we represent one of the attractors of the network. Active nodes are represented in black; inactive nodes are depicted in white. The attractor may be reached by controlling the state of the three nodes highlighted in red. We identified the seed set using the proposed algorithm; the seed set is $\mathcal{X} = \{(\mathrm{SLP}, \hat{\sigma}_{\mathrm{SLP}} = 0), (\mathrm{nWG}, \hat{\sigma}_{\mathrm{nWG}} = 1), (\mathrm{nHH}, \hat{\sigma}_{\mathrm{nHH}} = 1)\}$. (b) The 10 attractors of the network, and their corresponding minimal driving sets as identified by the proposed algorithm. Attractor 4 is the same as in panel a. Our predictions recover exactly the minimal driver sets for 7 of the 10 fixed points. We overstimate the size of the driver sets required to reach attractors 5, 9 and 10 by one node only (the PTC node). (c) Residual entropy as a function of size of the seed set for three selected seed sets leading to specific attractors. The labels of the attractors are the same as in panel b. The unconstrained greedy selection process finds attractor 4. As a term of comparison, we display also the curve corresponding to seed sets composed of randomly selected nodes' indices/states.

Integrating external information into a graph community detection algorithm to achieve high-quality communities

Ariel Berardino[1], Natalí B. Rasetto[1], Alejandro Schinder[1], and Ariel Chernomoretz[1,2]

[1]Fundación Instituto Leloir
[2]Physics Department, FCEN, University of Buenos Aires/IFIBA (CONICET)

November 8, 2022

Keywords — clustering analysis, community detection algorithm on weigthed graph, continuous process, single-cell/single-nuclei RNAseq technique

1 Introduction

In the field of single cell transcriptomics cells are explored analyzing the density distribution heterogeneity of ensembles of single-cell transcriptional profiles. In this context it is extremely important to identify high quality communities that would led to the discovery of new cell types or cell stages. A general approach to deal with this pattern recognition problem in such a high dimensional space is to focus on a low-dimensional manifold approximation captured by a mutual K-nearest neighbors (MKNN) graph. There are many unsupervised community detection algorithms in graphs that seek to group data-sets according to different figure of merit. However, the community recognition task is an ill-posed problem and different algorithms typically produce different partitions of the data. In this work we address this issue and introduced scBioMerging: a method that integrate external information to identify robust and biologically relevant communities in single-cell transcriptional landscapes.

2 Methods

We aimed to get a biologically meaningful similarity measure between assayed cells. We started from a gene expression matrix obtained in a single-cell RNAseq experiment and constructed a mutual k-mutual nearest neighbor graph (**graphX**) based on the correlations between cells in the Principal Component space. Then, we calculated the standardized transcriptional profile of each cell (Z_i) in graphX. At the same time, we identified over-represented Gene-Ontology Biological Processes (i.e. **external information**) and, for each cell, computed a biological enrichment profile that we used to embed the assayed cells in a kind of biological space. An MKNN graph was then constructed based on the correlation between cell enrichment profiles (**graphBP**). Finally, we computed a **biological process similarity matrix** using a topological measure of similarity from graphBP and used it to weighed the graphX edges. In this way a scalar field that captured the biological-similarity between linked pairs of nodes was incorporated into graphX.

The idea was then to used an heuristic similar to the one implemented in the Louvain algorithm. We started with a high resolution partition of nodes (obtained by applying a k-mean community detection algorithm in PCA space) and we considered the corresponding **community graph** (each node representing a cluster from graphX nodes). Accumulated biological similarity was considered to weigh self-loops and inter-community links in order to look for partitions that maximize the modularity by merging clusters that were biologically similar and produce a new enhanced partition (P_i). This process was repeated about 10 times for different k-means initializations. Finally, a partition (P_f) was generated by applying hierarchical clustering on the adjacency matrix calculated with a voting method that weighs the matrices of the different partitions P_i (**WEAC**)(Dong Huang, 2015).

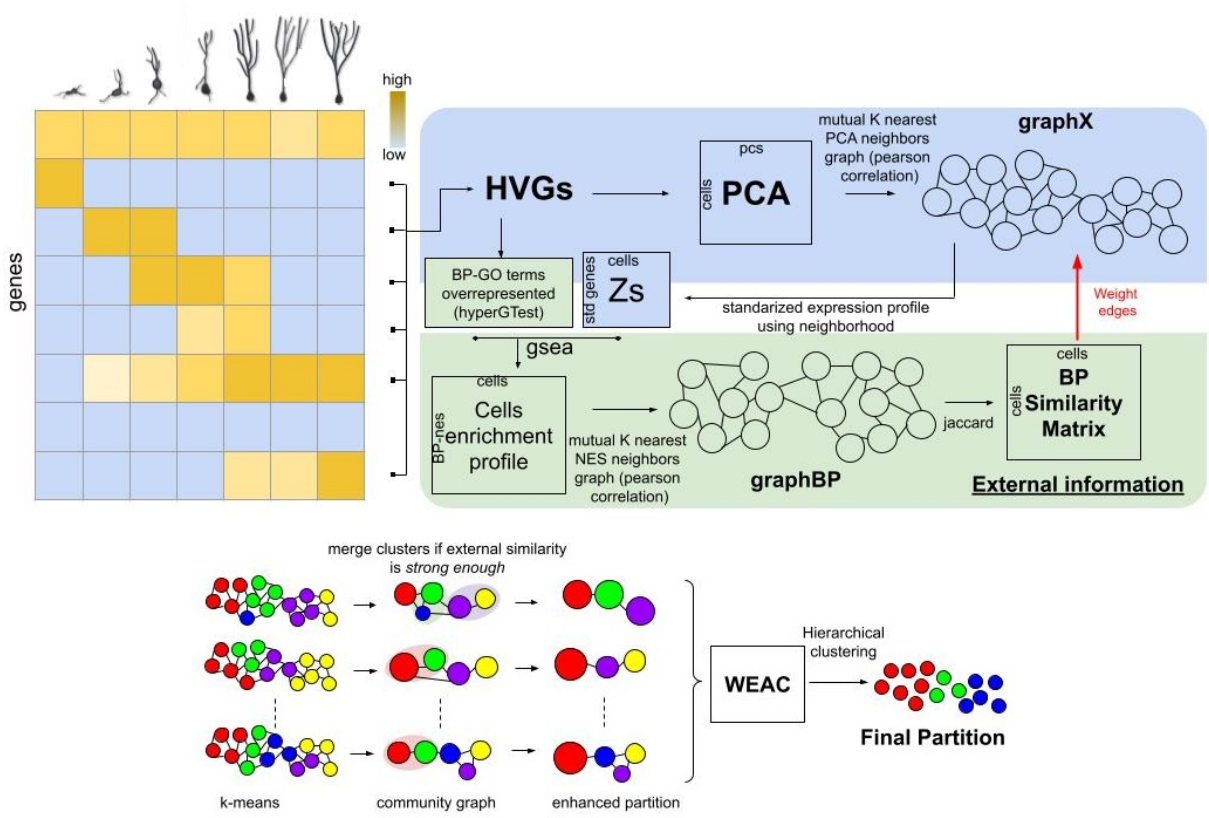


Figure 1: Scheme of scBioMerging's pipeline. We start from top left corner with the gene expression matrix, then we use the highly variable genes (HVGs) to compute de Principal Component Analysis Matrix. Using pearson correlation on the cells embedded in PCA space we built a mutual k-mutual nearest neighbor (MKNN) graph (graphX) on this expression space. We used this graph to calculate the standardized transcriptional profile of each cell in Zs matrix. We also used the HVGs to identify over-represented Gene-Ontology Biological Processes and then we computed a biological enrichment profile to embed the assayed cells in a kind of biological space. We used the same approach as in the expression space to construct a MKNN graph (graphBP) and then calculated a similarity matrix of biological processes for the cells. This matrix served as external information to be inyected in edges of graphX and capture the biological-similarity between linked pairs of nodes. Accumulated biological similarity was considered to weigh self-loops and inter-community links in order to look for partitions that maximize the modularity by merging clusters that were biologically similar and produce a new enhanced partition. This process was repeated about 10 times for different k-means initializations. Finally, a final partition was generated by applying hierarchical clustering on the adjacency matrix calculated with a voting method that weighs the matrices of the different partitions.

3 Results

Using scBioMerging on single-cell and single-nuclei RNAseq developmental datasets, we found clusters that were remarkably similar to those annotated by the authors of published papers (a.k.a "ground truth"). These clusters served as a solid starting point for: the identification of meaningful marker genes or the analysis of differential expression patterns between putative cell types or developmental stages.

The clusters provided by our method served as a solid starting point for: the identification of meaningful marker genes or the analysis of differential expression patterns between putative cell types or developmental stages.

4 Conclusions

We propose a novel and robust method that uses external information to generate a well defined partition on a continuous process. In particular, it can be used to identify biologically relevant cellular stages in a developmental dataset produced by single-cell/single-nuclei RNAseq techniques. We hope that it will help researchers in the analysis of this type of datasets and that they can find important cell stages that have a fundamental role in their developmental study.

References

C.-D. W. Dong Huang, Jian-Huang Lai, Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis, Neurocomputing 170 (2015) 240–250.

Integration of bio-medical information in a multimodal complex network for gene-disease prioritization.

Ingrid Heuer[1] and Ariel Chernomoretz[1,2]

[1]Physics Department, FCEN, University of Buenos Aires
[2]IFIBA (CONICET)

Keywords: gene-disease networks, data integration, disease-disease networks

One of the biggest challenges in current biomedical research is trying to bridge the gap between the different scales of organization that coexist in an organism, solving what is known as phenotype-genotype relationship. Ultimately, this translates into finding the molecular basis of different biological functions or pathologies and diseases. Identifying relationships between disease phenotypes and genetic alterations is essential to better understand disease etiology and to improve genome-based diagnostics. However, experimental methods designed to find these associations can be expensive and time consuming. To address this challenge, we took advantage of the vast amount of available biomedical data and integrated a multimodal complex network that could be used to prioritize novel gene-disease associations.

In this work, we construct and analyze a multimodal biomedical knowledge graph that contains data about gene-disease associations, complemented with protein-protein interactions, biological pathways , disease ontology relationships and natural language descriptions of the involved diseases. This multimodal network integrates quality resources such as DisGeNET [1], HIPPIE [2], PrimeKG [3], Reactome [4] and Signor [5] (table 1). The integrated network consists of 5 types of nodes organized in two layers: a disease layer and a gene/protein layer. The two layers are connected by gene-disease associations (fig. 1a). We found that 99% of nodes of our multimodal network were at less than three hops away from a node of the complementary layer.

We considered two disease ontologies in our network, MONDO [6] and UMLS[7], which we combined using vocabulary mapping. The integration of knowledge embedded in disease ontologies is a challenging task, since the definition of a unique disease is ambiguous and often inconsistent between databases. To address this issue, we incorporated BERT-group nodes in our network, which are disease concept groups obtained using the ClinicalBERT natural language processing model [3].

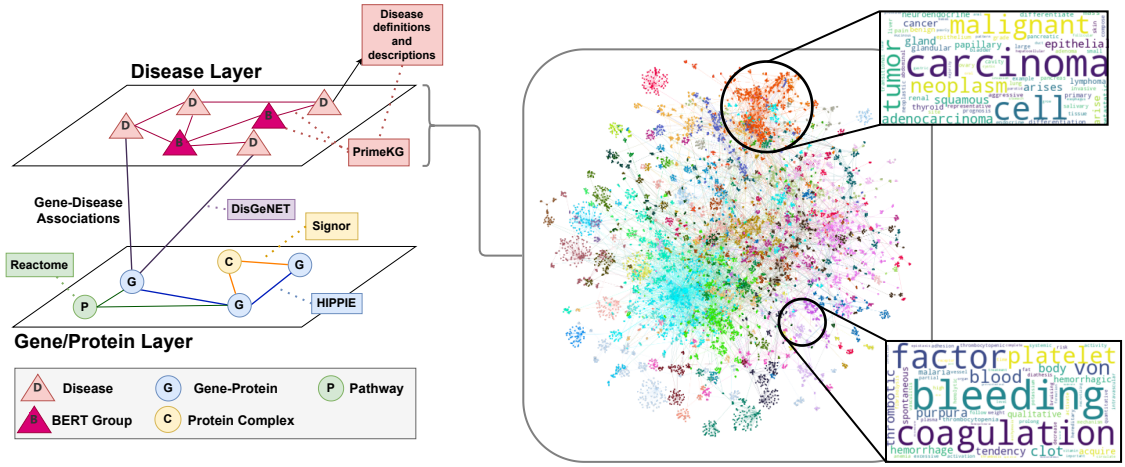
In order to probe the disease layer, we studied its mesoscale structure at two different resolutions considering two community detection algorithms: Infomap and Louvain (fig 1b). In each case we characterized the detected communities in terms of the homogeneity of their components considering two metrics. The first one was based on shared gene associations between diseases whereas the second one considered the semantic similarity of disease nodes inferred from a TF-IDF analysis of their natural language descriptions. To that end we used a measure of semantic specificity, $Spec$, based on the entropy of the TF-IDF distributions associated with descriptions of the components of a community:

$$Spec_j = 1 - H_j = 1 + \sum_{i=1}^{N} p_{i,j} \log p_{i,j} \tag{1}$$

where N is the number of terms in the network corpus and $p_{i,j}$ is the TF-IDF value of the term i in the community j. We compared the semantic specificity of these communities with a randomly generated control sample and found that the communities detected by both algorithms showed significant semantic specificity (fig 2).

We also studied the structural role of BERT-group nodes in the disease layer. Using set similarity metrics, we compared disease communities with groups of nodes that belong to BERT-groups, and found that communities tend to form around BERT nodes. To further understand the role of disease group nodes, we characterized them in terms of their participation coefficient and within-module degree [8]. We saw that BERT-groups tend to have a connective role within their communities and a non-zero participation coefficient, which indicates that they act as module connectors.

Overall, we were able to build a bio-medical network integrating more than 10000 diseases and 84000 high confidence gene-disease associations with protein-protein interaction and biological pathway information. In particular we analyzed different features of the disease layer and found that the observed connectivity patterns could provide a meaningful scaffold to implement message passing algorithms for link prediction and prioritization tasks.



(a) Multimodal network diagram

(b) Disease layer

Figure 1: (a) Diagram of the integrated multimodal network: The network is organized in two layers: a disease layer and a gene/protein layer. The disease layer contains disease nodes and BERT-group nodes. The gene/protein layer contains gene/protein, pathways and protein complex nodes.(b) Disease layer: Colors indicate communities detected with the Louvain algorithm. Wordcloud examples show the most relevant terms associated with a community, which were extracted from the descriptions of the diseases that belong to that community using a TF-IDF approach.

Node Type	Number	Source
Disease	15766	DisGeNET
BERT-Group	1067	PrimeKG
Gene/Protein	17363	DisGeNET-HIPPIE
Complex	422	Signor
Pathway	2020	Reactome
Total	36638	

Edge Type	Number	Source
Gene-Disease Association	84038	DisGeNET
Disease-Disease	17488	PrimeKG-MONDO
Protein-Protein Interaction	110062	HIPPIE
Pathway-Protein	42646	Reactome
Protein-forms-Complex	1888	Signor
Total	256122	

Table 1: Number of nodes and edges in the network, node/edge type and source database for each type.

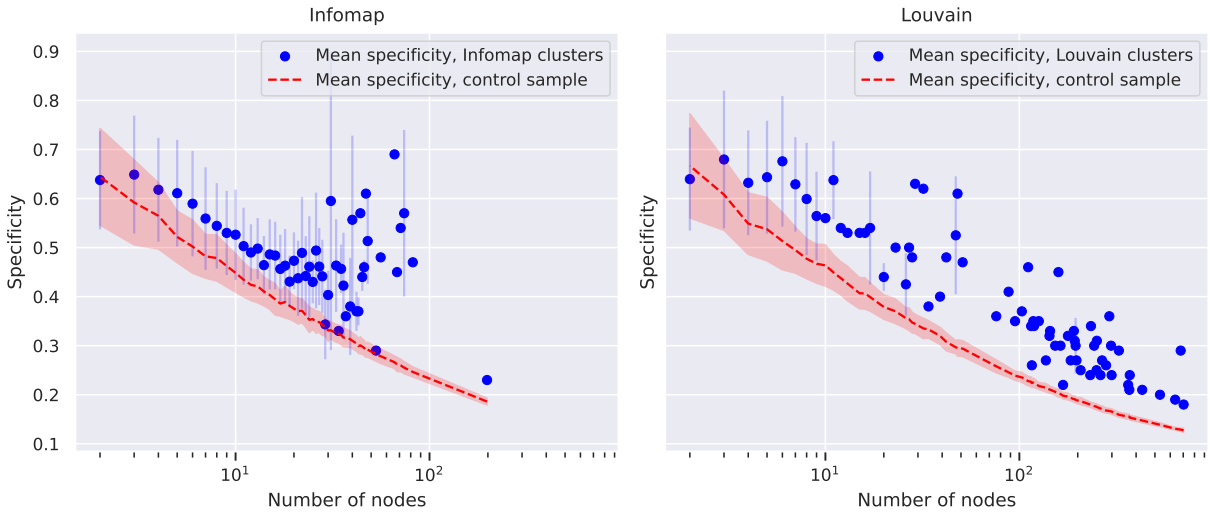


Figure 2: Semantic specificity of communities in the disease layer. Blue markers show the mean specificity of groups of communities of the same size. We compared this metric with a randomly generated sample of communities, shown in red. We found that the communities detected by both algorithms showed significant semantic specificity

References

[1] Janet Piñero et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019.

[2] Alanis-Lobato et al. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 10 2016.

[3] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *bioRxiv*, 2022.

[4] Marc Gillespie et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 11 2021.

[5] Luana Licata et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Research*, 48(D1):D504–D510, 10 2019.

[6] Christopher J. Mungall et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722, 11 2016.

[7] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, January 2004.

[8] Roger Guimerà and Luís A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, February 2005.

Lateralization properties in motor brain networks

Juliana Gonzalez-Astudillo[1], and Fabrizio De Vico Fallani[1]
[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

There is a direct relationship between functional specialization and the spatial organization of the human brain. This is not a random organization, on the contrary, it follows a precise order like proximity between complementary areas or functional symmetry across hemispheres [1]. A clear example is motor function, which principally involves the motor cortex, but still needs interactions with somatosensory areas for a proper preparation of the movement (see Fig. 1-A). Besides, it characterized for presenting a particular asymmetry in which each hemisphere is principally involved in controlling the contralateral side of the body [2, 3].

From this brain lateralization it emerges the qualitatively differentiation between within- and across-hemisphere interactions, that influence the strength of a region or node depending on how these contributions are conceived. By considering homotopic locations in the two hemispheres, lateralization can then be quantified using two separate metrics: segregation (σ_{ij}) and integration (ω_{ij}). The first measures the tendency for greater within-hemisphere interactions compared to between-hemisphere interactions

$$\sigma_{ij} = \frac{(LL_i+LC_i-LR_i)-(RR_j+RC_j-RL_j)}{(CL_k+CR_k+CC_k)}, \quad (1)$$

where each term represents the strength of a node in the homotopic pair of nodes i and j. In the differentiation between within- and across hemispheres edges, the capital letters respectively denotes the locations of node i and the nodes it establishes connections with (e.g. LR_i means that node i belongs to the left hemisphere and we consider the connections that link it to the right hemisphere nodes, see Fig. 1-B). Note that for the particular case of brain signals recorded with an EEG system, the electrodes placed in the midline sagittal plane (C_k) do not strictly belong to a hemisphere, so we consider them to normalize the metrics values.

Applying the same notation, integration seeks the contribution of contralateral connections, characterizing how the information flows across hemispheres. Then it is defined as the summed effect of within- and across-hemispheric interactions

$$\omega_{ij} = \frac{(LL_i+LC_i+LR_i)-(RR_j+RC_j+RL_j)}{(CL_k+CR_k+CC_k)}, \quad (2)$$

To prove the relevance of these metrics in characterizing lateralized cognitive process, we studied EEG signals from 140 subjects performing motor imagery of the right and left hand [4]. We estimated spectral coherence-based networks and we computed the previously described network lateralization metrics for each node or electrode.

We evaluated the presence of specific task-associated patterns for each metric by statistically compearing both motor conditions. We performed a t-test at the subject level and for each node, assuming a null hypothesis that the two means (σ_{ij} or ω_{ij}) were equal. We resumed the obtained results in Fig. 1, where for illustrative purposes, we show the mean t-values across subjects.

This analysis enabled us to identify the most discriminant electrodes. For both metrics engage a subset of nodes mostly located in the M1 cortex, but also the PMA, SMA and S1 areas also crucial in the planification and execution of a movement [5]. We observe that ω shows higher values, while σ also involves frontal areas, usually associated with attention and motor planning. These results show the neurophysiological plausibility of our proposed network approach. Moreover, they prove to be highly relevant features for decoding a MI mental task.

[1] SJ. Gotts, HJ. Jo, GL. Wallace, ZS. Saad, RW. Cox, and A. Martin, *Two distinct forms of functional lateralization in the human brain*, Proceedings of the National Academy of Sciences **36**, E3435-E3444 (2013).

[2] G. Pfurtscheller, and FH. Lopes Da Silva, *Event−related EEG/MEG synchronization and desynchronization: basic principles*, Clinical neurophysiology **110**, 1842-1857 (1999).

[3] T. Cattai, S. Colonnese, MC. Corsi, DS. Bassett, G. Scarano, and F. De Vico Fallani, *Phase/amplitude synchronization of brain signals during motor imagery BCI tasks*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1168-1177 (2021).

[4] V. Jayaram, and A. Barachant, *MOABB: trustworthy algorithm benchmarking for BCIs*, Journal of neural engineering **15**, 066011 (2018).

[5] S. Hétu, M. Grégoire, A. Saimpont, MP. Coll, F. Eugène, PE. Michon, and PL. Jackson, *The neural network of motor imagery: an ALE meta-analysis*, Neuroscience & Biobehavioral Reviews **37**, 930-949 (2013).

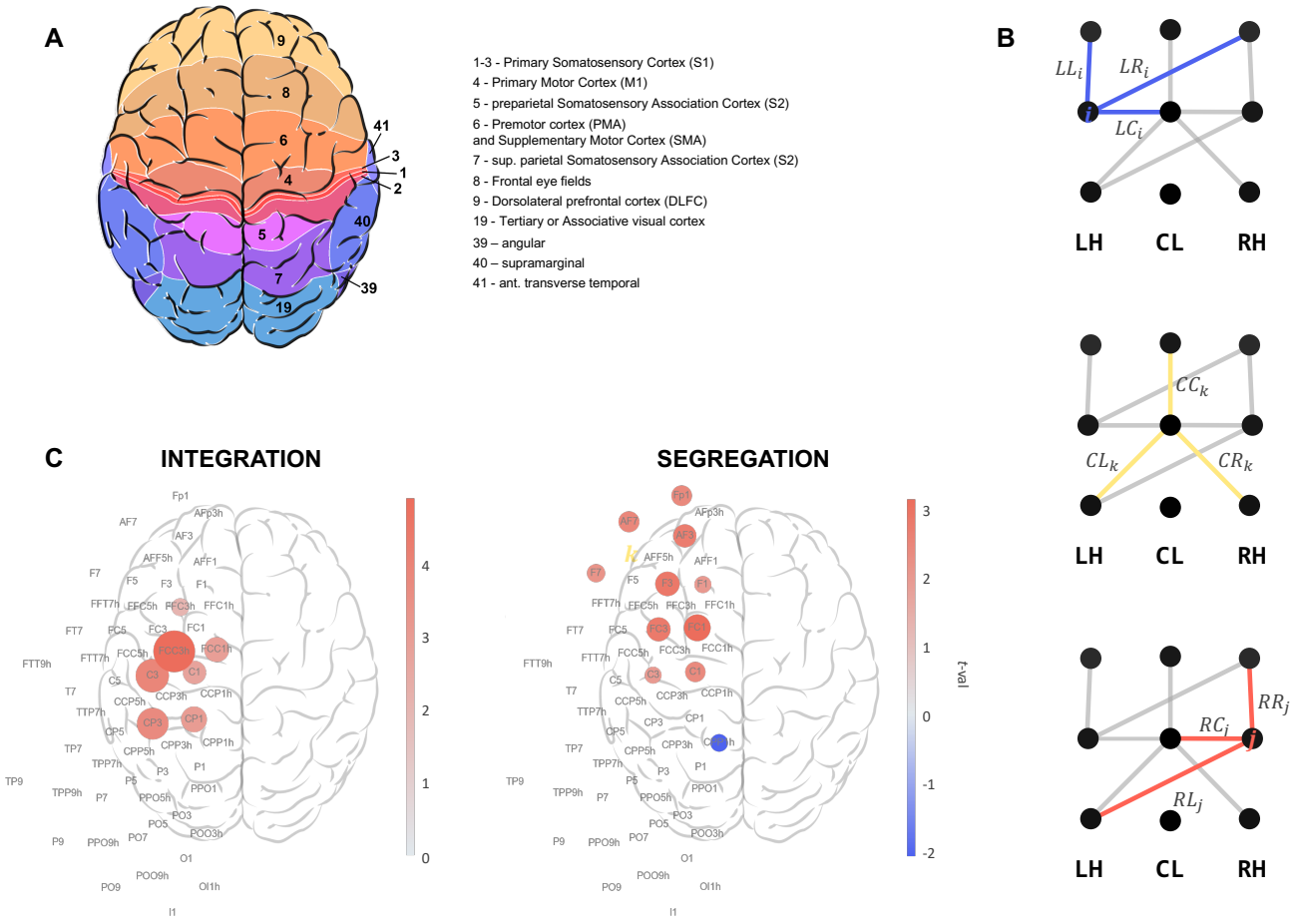


Fig. 1. A- The functional units of the cerebral hemispheres have been separated into what are called Brodmann areas. Motor cortex (M1) is area 4; the primary sensory cortex (S1) includes areas 3, 1, and 2. B- Within and inter-hemisphere connections. LH: left hemisphere, RH: right hemisphere and CL: central line. C- Group-averaged node-t-values between right and left MI mental states. By definition, lateralization metrics are anti-symmetric with respect to the hemispheres. For the sake of simplicity, only the left hemisphere is shown in here.

Migration Reframed? Multilingual analysis on the stance shift in Europe during the Ukrainian crisis.

Sergej Wildemann
L3S Research Center
Germany

Erick Elejalde
L3S Research Center
Germany

Keywords: migration, social media, stance detection, multilingual analysis

Abstract

The war in Ukraine created a large wave of refugees leaving the country. Four months after the 2022 invasion, the UN Refugee Agency (UNHCR) reported 5.1 million[1] Ukrainian refugees [1]. By September 2022, more than 4M people have registered for protection schemes in European countries, especially in Poland with more than 1.3M and Germany with more than 700k persons[2]. This makes the situation in sheer numbers comparable with the European refugee crisis of 2015-2018.

The conflict in Ukraine and the resulting migration received much attention in mass and social media. At the same time, there seems to be a more positive framing of migration, specifically toward refugees from Ukraine. Before the escalation of the conflict, media coverage and government policies on migration mainly focused on other groups coming from conflict zones, such as Syria, Ethiopia, or Afghanistan. However, they tend to be addressed in the context of economic and security threats [6, 3]. Attitudes towards migration are influenced by information from the press [8, 5] as well as by political agendas in the respective European countries. Media analysis has shown frequent negative terminology such as 'illegal', 'violence', 'terrorist' used in this context [7, 4, 3]. Moreover, news reports have linked certain crimes or socioeconomic issues (e.g., the rise of unemployment) to immigration [14, 2]. In contrast, in 2022, the humanitarian crisis at the EU border resulting from Ukrainians fleeing the war prompted a massive reaction of support by the western media and a great display of solidarity from the European public in particular. Several differences might have triggered this change, including (a) the intensive reporting about the war situation raises empathy (there was also a more "welcoming" culture during the war in Syria) and (b) the demographic difference in the migrant population (about 90% of the refugees from Ukraine are female[2]). Another significant difference seems to be (c) the cultural proximity of Ukraine to the EU compared to refugees from the Middle East [12].

We investigate whether the impression of a stance shift towards migration is substantiated by how the topic is reflected in online news and social media, thus linking the representation of the issue on the Web to its perception in society. The abundance of audience interactions on Twitter with the news provides a precious source of data for understanding users' engagement patterns and evolving opinions on sensitive topics [11, 10]. Starting from 5.5M Twitter posts published by 565 European news outlets in one year, beginning September 2021, plus replies from their audiences, we perform a multilingual analysis of migration-related media coverage and associated social media interaction in Europe. In particular, we focus on the five western and central European countries France, Germany,

[1] Until June 16th, there were 7.7 million border crossing from Ukraine since February, but 2.5 million crossed back to Ukraine in the same period [1].

[2] `https://data.unhcr.org/en/situations/ukraine`

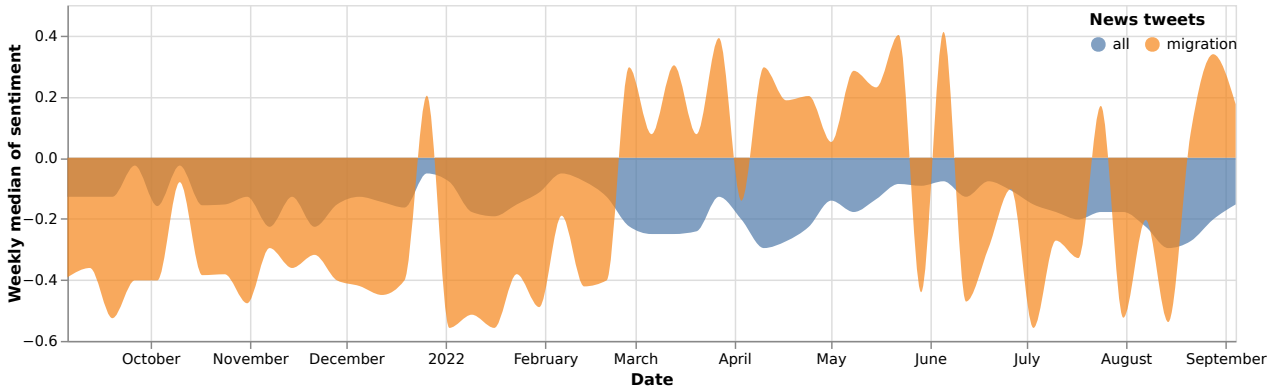


Figure 1: Median online sentiment over time in all news (blue) and migration-related news (orange). Poland 2021-2022.

Italy, Poland, and Spain. To characterize the change in the tone of the discussion, we contrast the language and related sentiment used by the mass media in their online coverage of refugee crises before and after February 2022 (see an example for Polish media in Figure 1). Moreover, we examine the reaction of the EU audience through their engagement on Twitter with related news. We use an original and effective methodology based on NLP and machine learning to learn the users' stances. The novelty of our approach to stance detection lies in its systematic multilingualism and in the context-dependent way the content is collected. In contrast to other approaches [9, 13] we do not rely on keywords or hashtags to select suitable responses to measure a user's targeted attitude toward migrants and refugees, but instead leverage the topics set by the news.

The results of our analysis show that there is a reframing of the discussion in news media. Comparing November 2021 and March 2022, this is evident in the change in terminology, e.g., from "migrant" to "refugee", often even accentuated by phrases such as "real refugees". Pre-invasion media discourse concerns seemed dominated by ethnocentrism or worries about the impact of migration on society. In terms of a change in public perception, however, the picture is more differentiated than expected. We can observe a noticeable stance shift in the positive direction for all countries at the beginning of the Ukraine invasion, starting in February 2022. This change can be seen most clearly in Poland, which is also the country most affected by the subsequent refugee movement. There, a lasting change in attitude is also evident, while the effect in the other countries is less stable and fades after about 3 to 4 months. In contrast, the impact of another recent crisis at the Belarus-EU border[3] in Poland is accompanied by a noticeably negative attitude both in the public and media coverage. Additionally, we find that the media sentiment leads the public stance with the one-week lag showing the highest F-statistic in a Granger causality test ($F = 26.11, p < .0001$).

Lastly, we look at the performance of cross-lingual stance classification. By training the model on all languages except the test language, we investigate the transferability of knowledge from richer datasets to other languages. Our experiments demonstrate that a model tuned for this task can sufficiently generalize to be applied to additional languages or countries.

Our multilingual stance detection method has proven very effective for the study performed. Generalizability to unseen languages can further extend the impact of our methodology by easing the comparison of the situation in multiple countries, which can yield further insights into societal processes such as migration. There are various directions for future sociology and technology research. Examples include a deeper analysis of the reasons for the differences between European countries, distinction in the interaction of individual media with their audience, and the analysis of narratives around migration.

[3] `https://www.iom.int/news/iom-and-unhcr-call-immediate-de-escalation-belarus-poland-border`

References

[1] UNHCR The UN Refugee Agency. 2022. *Operational Data Portal: Ukraine Refugee Situation.* UNHCR. Retrieved June 21, 2022 from `https://data.unhcr.org/en/situations/ukraine`

[2] Christine Benesch, Simon Loretz, David Stadelmann, and Tobias Thomas. 2019. Media coverage and immigration worries: Econometric evidence. *Journal of Economic Behavior & Organization* 160 (2019), 52–67.

[3] Glenda Cooper, Lindsey Blumell, and Mel Bunce. 2021. Beyond the 'refugee crisis': How the UK news media represent asylum seekers across national boundaries. *International Communication Gazette* 83, 3 (2021), 195–216.

[4] Mathieu Couttenier, Sophie Hatte, Mathias Thoenig, and Stephanos Vlachos. 2019. The logic of fear-populism and media coverage of immigrant crimes. *Available at SSRN 3328507* (2019).

[5] James Dennison and Lenka Dražanová. 2018. *Public attitudes on migration: rethinking how people perceive migration: an analysis of existing opinion polls in the Euro-Mediterranean region.* Technical Report. European University Institute.

[6] Jakob-Moritz Eberl and Sebastian Galyga. 2021. Mapping media coverage of migration within and into Europe. In *Media and Public Attitudes Toward Migration in Europe.* Routledge, 105–122.

[7] Jakob-Moritz Eberl, Christine E Meltzer, Tobias Heidenreich, Beatrice Herrero, Nora Theorin, Fabienne Lind, Rosa Berganza, Hajo G Boomgaarden, Christian Schemer, and Jesper Strömbäck. 2018. The European media discourse on immigration and its effects: A literature review. *Annals of the International Communication Association* 42, 3 (2018), 207–223.

[8] Joana Kosho. 2016. Media influence on public opinion attitudes toward the migration crisis. *International Journal of Scientific & Technology Research* 5, 5 (2016), 86–91.

[9] Sumeet Kumar, Ramon Villa Cox, Matthew Babcock, and Kathleen M Carley. 2021. A Weakly Supervised Approach for Classifying Stance in Twitter Replies. *arXiv preprint arXiv:2103.07098* (2021).

[10] Jeffrey Lazarus and Judd R Thornton. 2021. Bully pulpit? Twitter users' engagement with President Trump's tweets. *Social science computer review* 39, 5 (2021), 961–980.

[11] Alessandro Rovetta, Lucia Castaldo, et al. 2021. Influence of mass media on Italian web users during the COVID-19 pandemic: infodemiological analysis. *JMIRx med* 2, 4 (2021), e32233.

[12] Lenka Thérová. 2022. Anti-Immigration Attitudes in Contemporary Polish Society: A Story of Double Standards? *Nationalities Papers* (2022), 1–16.

[13] Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691* (2020).

[14] Iris Wigger. 2019. Anti-Muslim racism and the racialisation of sexual violence:'intersectional stereotyping'in mass media representations of male Muslim migrants in Germany. *Culture and religion* 20, 3 (2019), 248–271.

Modeling the spatiotemporal epidemic spreading of multiple virus strains

Wesley Cota[a,b], Pâmela S. Andrade[c], Raissa H. A. Eliodoro[b], Franciane M. de Oliveira[d], Pedro S. Peixoto[e], Nuno R. Faria[b,f,g], Ester C. Sabino[b], Carlos M. C. B. Fortaleza[a]

[a]*Faculdade de Medicina de Botucatu Universidade Estadual Paulista Brazil*
[b]*Instituto de Medicina Tropical Universidade de Sao Paulo Brazil*
[c]*Faculdade de Saude Publica Universidade de Sao Paulo Brazil*
[d]*Faculdade de Medicina Universidade de Sao Paulo Brazil*
[e]*Instituto de Matematica e Estatistica Universidade de Sao Paulo Brazil*
[f]*Departamento de Molestias Infecciosas e Parasitarias Universidade de Sao Paulo Brazil*
[g]*Department of Infectious Disease Epidemiology and MRC Centre for Global Infectious Disease Analysis Jameel Institute School of Public Health Imperial College London London UK*

Keywords: epidemic spreading, metapopulation model, human mobility, SARS-CoV-2 variants

The mathematical modeling of epidemic diseases allows us to investigate the role of fundamental aspects in the spreading of infectious diseases. Compartimental models have been used with great success to describe different diseases and their aspects [1]. Network science has contributed with its convenient framework to represent the complex interactions of individuals and populations. The recent availability of large datasets, such as about human mobility and social behavior, has improved the accuracy of predictions achieved with epidemic modeling. In particular, the data-driven approach has been adopted by building analytically tractable models, but at the same time use explicit real data [2]. For example, origin-destination (OD) matrices are obtained by capturing the daily mobility of the population at different levels, from global with the airport networks, to the local scale, measuring the urban rhythms inside a country, state, or municipality. The mobility of spatially separated subpopulations can be represented by metapopulations, in which subpopulations are placed in different patches as nodes of a directed network whose interactions or edges quantify the flow of individuals from a patch i to j by an OD matrix W_{ij}. These approaches were successful in the context of COVID-19 such as in measuring the effects of non-pharmaceutical intervetions [3] and the outbreak diversity across different geographical scales [4].

In this work, we investigate a metapopulation modeling for the spread of multiple strains of the same virus, such as the SARS-CoV-2 variants. We assume that the population is distributed in Ω patches, each patch i containing a subset of n_i individuals, while the flow of individuals is governed by a normalized OD matrix $R_{ij} = W_{ij}/\sum_l W_{il}$ in a movement-interaction-return (MIR) model [2]. We assume a susceptible-exposed-infected-recovered compartmental dynamics in which individuals can be susceptible and infected by different strains, one at a time. The number of individuals in a given compartment $Z = \{S, E, I, R\}$, patch i, susceptible state ν and last infection state σ is represented by $Z_i^{\nu,\sigma}$. For three different strains, for example, we can define $\sigma = \{\sigma_1, \sigma_2, \sigma_3\}$ and $\nu = (\nu_1, \nu_2, \nu_3)$ in which $\nu_i = 1$ when the individual is not immune for the variant σ_i, and 0 otherwise, $i \in [1, 3]$. The model and rates are schematically shown in Fig. 1(a). Dynamical equations are written using a Microscopic Markov Chain Approach (MMCA) approach [2, 3]. We find multiple waves of infection governed by the existence of different variants, shown in Fig. 1(b,top) for three patches, and measure the impact of seeding a new variant in

different patches of the networks, measured by the number of exposed and infected individuals by each strain in each patch, Fig. 1(b,bottom). We compare the results with real data of the prevalence of each variant in the largest municipality of Brazil, Sao Paulo, for 10 months of 2021, sampled in different districts of the city with geo-referenced data. The empirical results are shown in Fig. 1(c), showing the replacement of one variant to another as in the simulated results. The perspective of this work is to calibrate the model together with a real OD network within the municipality via mobile geolocation data [5] and time-evolving rates as a function of vaccination.

W.C. acknowledges the grant #2021/11953-5, Sao Paulo Research Foundation (FAPESP).

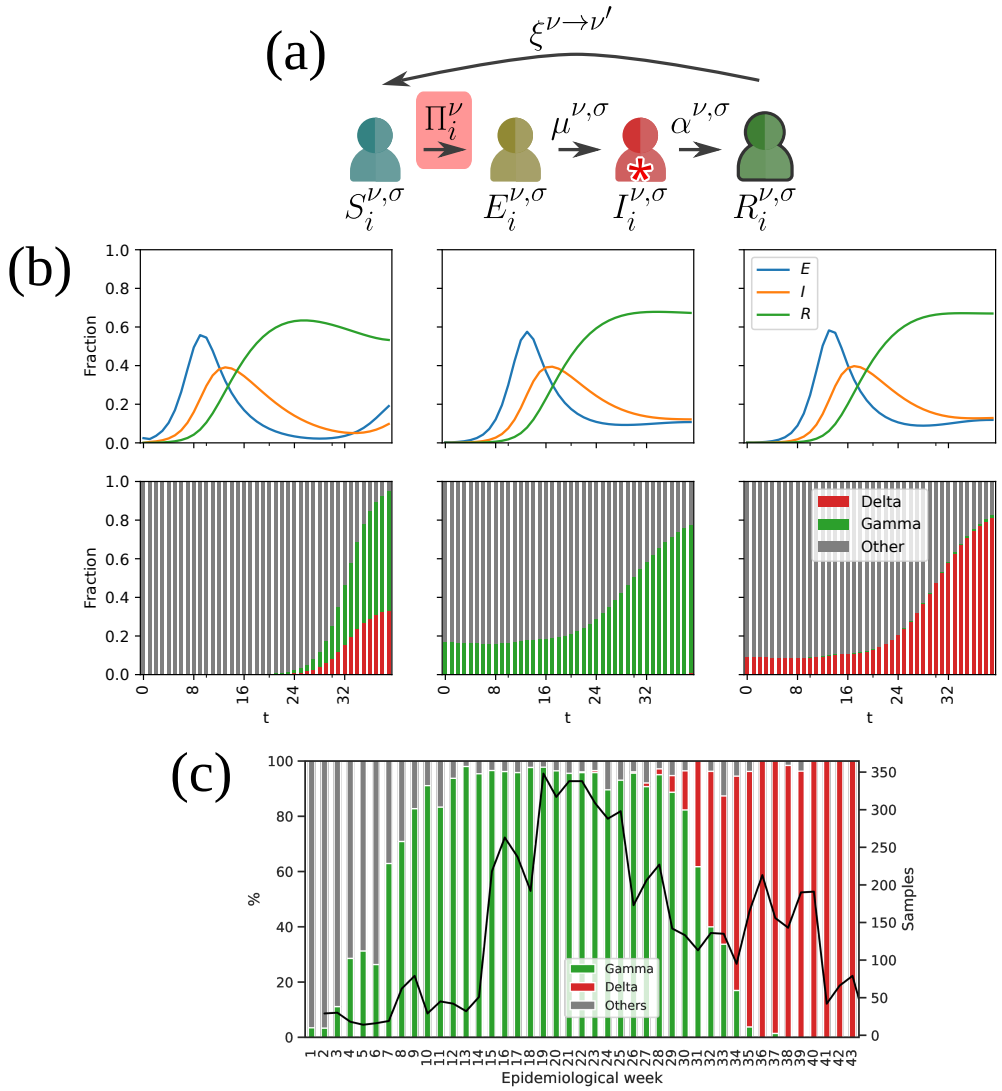


Figure 1: (a) Schematic representation of the model, its compartments and transition rates. Susceptible individuals are infected by a probability Π_i^ν, becoming an exposed individual $E_i^{\nu,\sigma'}$ with probability $\Pi_i^{\sigma'}$, that takes into account all infectious individuals with the strain σ' and the susceptibility ν currently in a patch i (residents or visitors), while recovered individuals become again susceptible to a new strain with probability $\xi^{\nu\to\nu'}$ spontaneously. (b) Multiple waves of infection (top) and prevalence of the variants in exposed and infected individuals (bottom) for three patches of a random geometric network with $\Omega = 50$ patches with subpopulations of size n_i given uniformly from 100 to 1000, and out-going flow $\sum_j R_{ij}$ sampled uniformly from 0.1 to 0.4. (d) Real data of the prevalence of distinct strains of SARS-CoV-2 in the municipality of Sao Paulo, Brazil.

References

[1] Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, Princeton, NJ, September 2011.

[2] David Soriano-Paños, Wesley Cota, Silvio C Ferreira, Gourab Ghoshal, Alex Arenas, and Jesús Gómez-Gardeñes. Modeling communicable diseases, human mobility, and epidemics: A review. *Annalen der Physik*, page 2100482, 2022.

[3] Alex Arenas, Wesley Cota, Jesús Gómez-Gardeñes, Sergio Gómez, Clara Granell, Joan T Matamalas, David Soriano-Paños, and Benjamin Steinegger. Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions. *Physical Review X*, 10(4):041055, 2020.

[4] Guilherme S Costa, Wesley Cota, and Silvio C Ferreira. Outbreak diversity in epidemic waves propagating through distinct geographical scales. *Phys. Rev. Research 2, 043306*, 2020.

[5] Pedro S. Peixoto, Diego Marcondes, Cláudia Peixoto, and Sérgio M. Oliva. Modeling future spread of infections via mobile geolocation data and population dynamics. an application to covid-19 in brazil. *PLOS ONE*, 15(7):1–23, 07 2020.

Modelling how social network algorithms can influence opinion polarization

Henrique Ferraz de Arruda [1], Felipe Maciel Cardoso[2], Guilherme Ferraz de Arruda [1], Alexis R. Hernández[3], Luciano da F. Costa[4], and Yamir Moreno[1,2,5]

[1]CENTAI Institute, Turin 10138, Italy
[2]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Spain
[3]Institute of Physics, Rio de Janeiro Federal University, Rio de Janeiro, RJ, Brazil
[4]São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, Brazil
[5]Department of Theoretical Physics, University of Zaragoza, 50018 Zaragoza, Spain
*h.f.arruda@gmail.com

Due to the considerable online social networks and their impact on society, scholars have been studying the dynamics behind these processes (1; 2; 3). Here, we proposed a novel approach to simulate online discussions on social networks, in which users and their friendship as the nodes and edges of a network. Our opinion model is based on the idea of adding information from external sources and how users and social network algorithms handle the information. The first step of this process represents a piece of news obtained outside the online social network, simulated as a random number. Next, two separate steps are associated with this news piece: *post transmission* and the *post distribution*. The *post transmission* models the willingness of a user to share the information, creating a post in the social network. To do so, we used different probability functions based on the difference between the post and the user's opinion. Among the possibilities of this function, we considered a function based on the cosine-squared, leading the users to post information they strongly agree with or disagree with. The latter represents the scenarios in which users react to the news. The social network algorithm must define the users who receive the information if the post is transmitted in a real social network. In order to simulate this step, we modeled the *post distribution*. More specifically, with the basis of the difference between the opinions of the posting user and their neighbors, another probability function is defined. Again, several different *post distributions* were tested. Finally, if the users receive posts that disagree with their opinions, there is a chance of breaking the friendship, which is modeled as a third probability function. If the friendship is broken, the respective edge is rewired. Several variations of *post transmission* and the *post distribution* were tested, as well as different network structures. Fig. 1 illustrates one step of the opinion dynamic and four possible results. Our dynamic converged into many different scenarios, including opinion consensus, polarization, and the formation of echo chambers. For various dynamic configurations, friendship rewiring can help promote echo chamber formation. However, even without allowing friendship rewiring, this effect can also occur for specific networks with well-defined community structures. Finally, we compare the results with real social networks. We show that the outcomes of our model are similar to the real scenario in terms of polarization and echo chamber formation. Our outcomes suggest that the social network algorithm can be important to mitigating or promoting opinion polarization. The work described in this abstract has been published in (4), and the source codes can be found at `https://github.com/hfarruda/OpinionPolarization`.

References

[1] C. Castellano, S. Fortunato, and V. Loreto, Reviews of modern physics **81**, 591 (2009).
[2] J. Lorenz, International Journal of Modern Physics C **18**, 1819 (2007).
[3] K. Sznajd-Weron, J. Sznajd, and T. Weron, Physica A: Statistical Mechanics and its Applications **565**, 125537 (2021).
[4] H. F. de Arruda, F. M. Cardoso, G. F. de Arruda, A. R. Hernández, L. da Fontoura Costa, and Y. Moreno, Information Sciences **588**, 265 (2022).

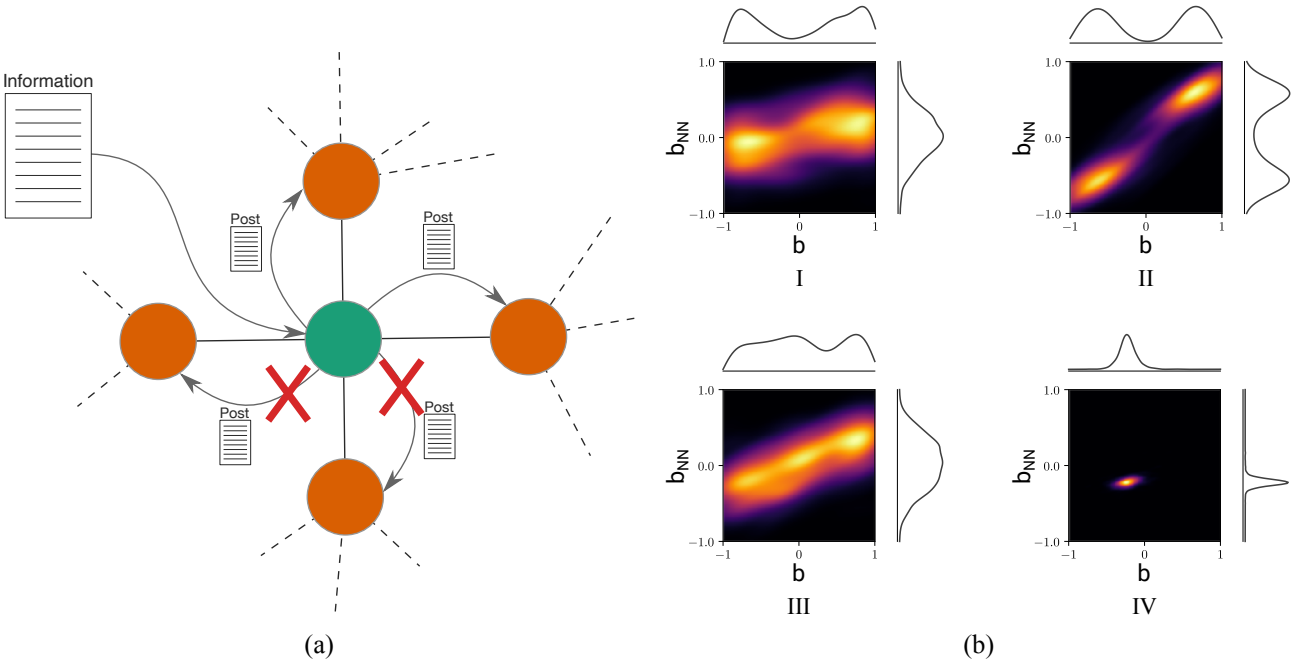


Figure 1: The scheme of the proposed model is shown in panel (a), in which a randomly selected user (green node) posts a piece of information, and two orange nodes receive the post. Panel (b) shows outcomes obtained from the execution of our dynamics, which illustrates some possibilities of the resultant opinion distributions. These plots represent the opinion assigned to nodes, b, against the average opinions of their neighbors (b_{NN}). Heatmap I represents the cases in which the opinions are polarized, but there is no echo chamber formation. This is evidenced by the fact that values of b_{NN} tend to be close to the average. In heatmap II the opinions are also polarized, but the values of b_{NN} tend to be closer to 1 or -1, representing the echo-chamber formation. In heatmap III, there is no polarization nor consensus, but extreme opinions exist. Another possible scenario is when the consensus is reached, illustrated in heatmap IV.

Natural language processing for understanding classroom dynamics

Bernardo García Bulle Bueno[1], Salome Aguilar Llanes[2], Tobin South[1], Alex 'Sandy' Pentland[1], and Esteban Moro[1,3]

[1]**Connection Science, MIT. Cambridge, MA**
[2]**Economics, MIT. Cambridge, MA**
[3]**Universidad Carlos III, Madrid**

ABSTRACT

Understanding the mechanisms of effective education is of paramount importance to society, and an area of research in need of more quantitative analysis. Much of the effectiveness of education settings lies in the nuance of the two way communication in the learning processes, and insights into how this communication meditates learning is hidden in natural language text. To understand more about the process, we performed an RCT on a tutoring program where university students teach Math to groups of up to five elementary school kids in 2021. We randomized both the tutor-group matching, as well as the focus of the class: just on Math or both Math and socioemotional learning. Using recordings of the sessions, we asked whether latent variables obtained from the scripts of the lessons could reveal (a) the mean increase in Math scores of the students in the group before and after participating in the program, and (b) the treatment that was assigned. We find that latent features from the script predict those metrics with 12 and 25% r-squared respectively. Math improvement was better predicted using latent variables from earlier layers of an encoder, while the opposite was true for the content. This suggests that identifying whether socioemotional content was part of the class was more related to the structure or meaning of the sentences, while Math improvement was more related to the words used.

Keywords: RCT, Education, NLP

1 INTRODUCTION

We propose to use tools from natural language processing to analyse the dynamics of learning within the classroom environment. To achieve this, we will use a new dataset collected through a tutoring RCT in Mexico comprising several thousands of hours of Math tutoring sessions with associated information on student and teacher outcomes in a Math standardized test implemented before and after tutoring session began. The tutoring connected groups of 5 elementary school kids to each tutor, and kids could be assigned to just Math content, or Math+socioemotional learning content. All the sessions were recorded and then analyzed by extracting latent variables from a Spanish language encoder. Automating the task of "observing" classroom dynamics and teaching style with a consistent toolkit allows us to understand how the detailed dynamics of classrooms affect outcomes at scale.

1.1 Related work

On the side of classroom observation the project conducted by the gates foundation "Measures of Effective Teaching" has a similar setting to ours as in that case classes were recorded and recordings were analyzed by observers using a standard set of questions that helped build an index of the measure. Students were randomly assigned to teachers within each school. The researchers observed that teachers that initially got a higher score of the measure were likely to get it again even if the students were changed to a new set of random students within the school [3]. There have been efforts to develop an internationally valid instrument to measure teacher effectiveness [5].

There are several examples of the use of Natural Language Processing (NLP) tools in education for instance the "E-rater" developed by ETS or Criterion to predict essay scores. Other existing tools such as "Text Adaptor" that help teachers develop text adaptations for their students. [2]. NLP can also aid in e-learning. For instance "Language Muse" that can give linguistic feedback to students. NLP can help in

the analysis of linguistic errors and aid as a tool to the teacher [1]. There also have been applications of NLP trying to replicate one-to-one human tutoring interactions. As well as to process web material to personalize instructional materials [4].

1.2 Data

In this work we analyze whether the script of a class can be used to predict either the Math improvement of the kids in that class, or whether the class covered only Math topics, or Math along socioemotional learning activities as well. To do it we passed the recordings of classes through an automatized speech-to-text software which yielded a text file containing the dialogue that was said in the class in a written form. Due to computational and budget constraints, only classes for 45 groups were analyzed in the following section.

2 REPRESENTATION LEARNING OF LATENT VARIABLES

To obtain latent variables from the scripts, we used an existing encoder architecture trained to understand Spanish. This approach is better than training a model on our own given our dataset is small compared to one used to train a language model. For this we used nvidia's Spanish English translator (24x6) with 24 encoding layers and 6 decoding layers. The architecture is a transformer one and includes key, query and value representations in the tensor. We only focused on the encoder part as that's the part that provides us with representations of the Spanish language.

We considered the intermediate product between layers of the encoder to be our representation. For each layer we analyzed, we extracted five variables: the mean of first n vectors (corresponding to n words), the last word vector (last word), and the 3 vectors corresponding to query, key and value, separately.

2.1 Relationship to outcome variables

We decided to use two outcome variables per group: Average improvement of kids in Math tests, and Assignment to emotional+Math tutoring (compared to just Math). Then we did two regressions explaining each of those outcome variables (y) with the latent variables we found (x). We use Lasso regularized OLS in both cases. To evaluate the performance of each we did k-fold out-of-sample predictions for each datapoint, and measured performance in terms of r-squared.

3 RESULTS

For this section we obtained latent variables from an the encoder of a Spanish-English translator as described in 2. We tested features from five parameters: mean word embedding, last word embedding, and query, key and value features from the transformers. Moreover, we also had the option to compare the results across different layers. The results are plotted in figure 1.

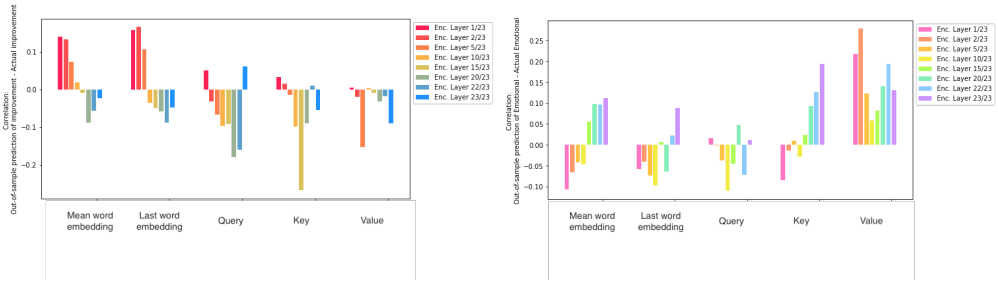


Figure 1. Results for different layers in terms of correlation between ground truth and out of sample k-fold predictions. The bar height is the median correlation across different regularization parameters.

We obtain very optimistic results in the representation learning. Curiously, for Math outcomes, the best results are found using early layer word encodings. As in other domains, the deeper intermediate layers are the worst, with an uptick in performance from the last layer. For the emotional treatment layer, the words are not as effective, and only are in the last layers. However, the value layer is surprisingly effective at capturing the emotional treatment class.

This result shows that encoded in the sentences of the class we have information which is related to the success and format of the class. The next step in our case was to be able to interpret the factors correlated with each of the outcome variables.

To be able to do it we used the original transformer model. We took a random subset of all sentences in all classes. Then for each sentence we obtained the values of the layer which we want to interpret. To do it, we generated predictions for the sentences and sorted them by the outcome variable to see how they differed.

In the case of the emotional treatment in the class, we found that longer sentences were associated with those classes. That is coherent with previous findings we observed which were that students felt more confident to talk in class in that treatment. This opens the door to being able to characterise a class by the emotional content of it and the bonding which happened in it.

Unfortunately, while we do see predictive power for characterising classes with more Math improvement, the methods we tried did not yield any results which we could qualitatively interpret with clarity, for which they are omitted.

4 DISCUSSION & CONCLUSION

This analysis has leveraged NLP to advance our knowledge of how education happens within the classroom. We find that despite the limitations of current pretrained models on Mexican Spanish, the use of multilingual models makes transcription effective and mostly robust. Utilising a learned representation space from pretrained translation models allowed for strong predictive performance of education outcomes with limited explainability. To address this we identify characteristic sentences that match high importance characteristics of the representation space and qualitatively identify that emotion and human connection in the educational context is a strong determinate of learning outcomes.

Our approach has the potential to help us understand how kids learn and can be used to help teachers improve their skills and learn from others. In future work, we propose to link more outcome variables to features encoded in language models. One important area of improvement we identified was our strategy to interpret the variables, as it was purely qualitative and relied on small samples of sentences. We hope to improve this by using automatic text generation techniques, and other visualisation techniques which can shed light on which exact components of the sentence resulted in higher perceived emotional content in the class.

REFERENCES

[1] Khaled M Alhawiti. “Natural language processing and its use in education”. In: *Computer Science Department, Faculty of Computers and Information technology, Tabuk University, Tabuk, Saudi Arabia* (2014).

[2] Jill Burstein. “Opportunities for natural language processing research in education”. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2009, pp. 6–27.

[3] Thomas J Kane et al. “Have we identified effective teachers? Validating measures of effective teaching using random assignment”. In: *Research Paper. MET Project. Bill & Melinda Gates Foundation*. Citeseer. 2013.

[4] Diane Litman. “Natural language processing for enhancing teaching and learning”. In: *Thirtieth AAAI conference on artificial intelligence*. 2016.

[5] Charles Teddlie et al. “The international system for teacher observation and feedback: Evolution of an international study of teacher effectiveness constructs”. In: *Educational research and evaluation* 12.6 (2006), pp. 561–582.

News-sharing on Twitter reveals emergent fragmentation of media agenda and persistent polarization

Tomas Cicchini[1,3], Sofia M. del Pozo[1,2], Enzo Tagliazucchi[1,2], and Pablo Balenzuela[1,2]

[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Buenos Aires, Argentina.

[2]Instituto de Física de Buenos Aires (IFIBA), CONICET. Buenos Aires, Argentina.

[3]Instituto del Cálculo (IC), UBA-CONICET. Buenos Aires, Argentina.

Abstract

News sharing on social networks reveals how information disseminates among users. In this work, we used bipartite news-user networks to study the news sharing behavior of main Argentinian media outlets in Twitter, in order to understand the role of political polarization in the emergence of high affinity groups with respect to news sharing. The behavior of users resulted in well-differentiated communities of news articles identified by a unique distribution of media outlets. In particular, the structure of these communities revealed the dominant ideological polarization in Argentina. We also found that users formed two groups identified by their consumption of media outlets, which also displayed a bias towards the two main parties that dominate the political life in Argentina.

Keywords: Social network analysis, News consumption, Natural language processing

Introduction

The mass media play a preponderant role in the formation of public opinion [1, 2]. Nowadays, many people are exposed to news and share them through social media [3, 4]. Therefore, understanding the way in which information circulates through them plays a fundamental role in the process of opinion formation.

Information flow is strongly influenced by how users connect among them in different social media. This connectivity [5] arises with formation of ties based on affinity, group membership or trust in influential individuals or organizations, among other causes [6]. As these ties emerge, social networks become clustered, leading to constraints on information flow.

The emergence of highly connected groups of individuals is a topological feature that repeatedly arises in studies of social networks in relation to discussions around specific topics. They have been observed in networks defined by preferential message propagation (*retweet networks*, in the case of Twitter) [7] , as well as in networks of followers [8]. These groups reflect the clustering of individuals based on different measures of similarity among users [8] creating homogeneous communities that are frequently known as *echo chambers* [9]. In these works, the reported polarization phenomena refers to groups which extreme their opinions on discussions around a specific topic (gun control, vaccination, etc.). However, topics are rarely discussed in isolation [10] and the phenomenon of issue alignment phenomena plays a key role in polarization in the political process, leading to antagonistic ideological states.

In this work we investigate which are the key features leading to the emergence of well defined groups in the process of news sharing of the main media outlets in Argentina. The main hypothesis of this work is that

the way in which users share news on a social networks, such as Twitter, is mediated by personal preferences and ideological affinity in such a way that it is possible to detect emerging groups as a consequence of these interactions.

Data & Methods

Twitter users sharing links to media content were selected in order to analyze the emergence of structures from their complex pattern interactions, following a similar approach than in [3]. We focused our analysis on two different periods: from August 29th to September 30th 2019 and from June 4th to July 4th 2020.

With the aim of investigate the leading of the emergence of well defined groups, the complex pattern of news shared by multiple users can be mapped onto bipartite networks following the procedure sketched in [3].

Bipartite networks have two different classes of nodes and can be projected into news and user layers. Connections in the news projection indicate co-consumption across users, while the user projection describes users connected by news in common. These two projections were analyzed using community detection algorithms, as well as external metrics of the news were taking account, such as the semantic content and the media outlet they belong to. Moreover, we use topic detection and sentimental analysis in order to understand the structure of the news projection. In Figure 1 the results of the news networks analysis are shown.

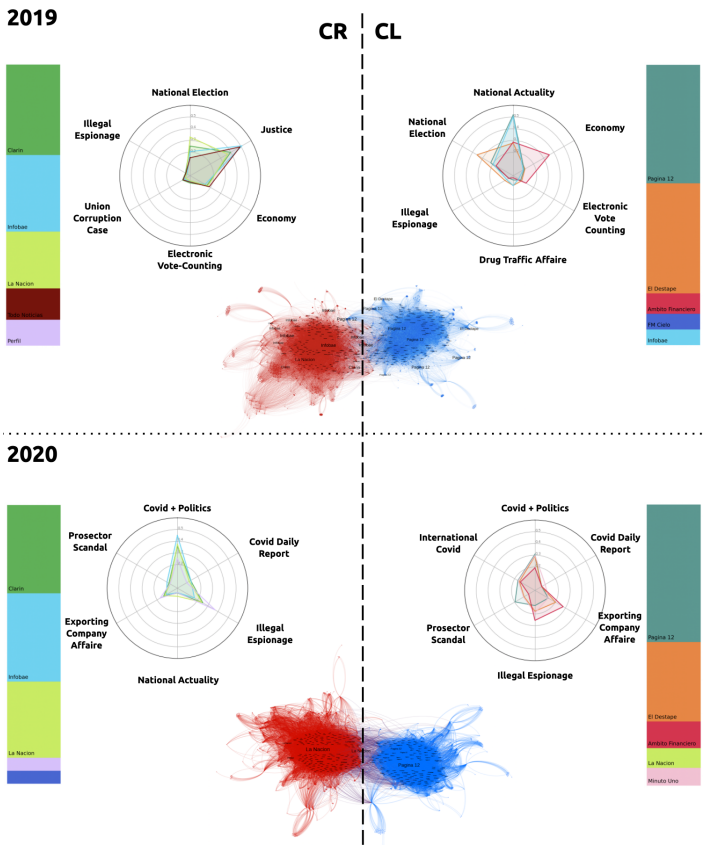


Figure 1: *Media outlets distributions and topic decomposition for the 2019 and 2020 two main communities. The stacked bars represent the media outlet distribution, while the radar plots display the media agenda. The agenda of each outlet is indicated with lines colored with the same color as in the stacked bar.*

News Networks

When detecting communities in the news networks we find a qualitatively correlation between the news media and the structure of communities. To quantify this, each community was described by a vector C_i^m, with each component associated with the amount of news from the media outlet m in the community i. With this, we calculate the cosine similarities between the main community vectors of the same year, and also from one year

to the other. Thus, we observe that there are subgroups of communities associated with certain media groups. In particular, a subgroup associated with center-left media and another, with center-right media. Then, we added the study of the sentiment bias around certain figures of national politics, with the aim of seeing if the Argentine political polarization was reflected in this community structure. These results confirmed that the center-left media set in both periods showed opposite behavior to that of the center-right. Further details can be found in [11].

User Networks

We also detect communities in the user network, finding a strong relationship between the media of news shared by users and the community to which they belong. Identifying users with a vector of media consumption m^i, we calculate the similarity between users and the average of these vectors of each community. We again find sets of communities more similar to each other, where the same media subgroups appear as before. Further details can be found in [11].

Discussion and Further Analysis

In this work we study if news sharing of Argentinian newspapers in social media produces the emergence of homogeneous groups in terms of media consumption habits. Our main contribution is the detection of echo chambers in bipartite user-news networks and their identification in terms of consumption patterns of media outlets associated with the underlying ideological leaning patterns in Argentinean political life.

Our results contribute to the characterization of echo chambers in terms of vector-media consumption and the use of sentiment bias to infer the leaning of media outlets. They also shed light on the process in which the political polarization in Argentina constrains the exposure to media content in social media.

References

[1] M. E. McCombs and D. L. Shaw. Public opinion quarterly. *Public Opinion Quarterly 36(2):176-187*, 1972.

[2] L. Guo and M.E. McCombs. *The Power of Information Networks: New Directions for Agenda Setting*. Routledge studies in global information, politics and society. Routledge, 2016.

[3] Iain S. Weaver, Hywel Williams, Iulia Cioroianu, Lorien Jasney, Travis Coan, and Susan Banducci. Communities of online news exposure during the uk general election 2015. *Online Social Networks and Media*, 10-11:18–30, 2019.

[4] Adrian Rauchfleisch, Daniel Vogler, and Mark Eisenegger. Transnational news sharing on social media: Measuring and analysing twitter news media repertoires of domestic and foreign audience communities. *Digital Journalism*, 8(9):1206–1230, 2020.

[5] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[6] Bin Gu, Prabhudev Konana, Rajagopal Raghunathan, and Hsuanwei Michelle Chen. Research note—the allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, 25(3):604–617, 2014.

[7] Carolina Becatti, Guido Caldarelli, Renaud Lambiotte, and Fabio Saracco. Extracting significant signal of news consumption from social networks: the case of twitter in italian political elections. *Palgrave Communications*, 5(1):91, Aug 2019.

[8] Cinelli, M., Morales, G., Galeazzi, A., Quattrociocchi, W. & Michele Starnini The echo chamber effect on social media. *Proceedings Of The National Academy Of Sciences.* **118**, e2023301118 (2021), https://www.pnas.org/doi/abs/10.1073/pnas.2023301118

[9] Kathleen Jamieson and Joseph Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press; First Edition (July 22, 2008), 01 2008.

[10] Baumann, F., Lorenz-Spreen, P., Sokolov, I. & Starnini, M. Emergence of Polarized Ideological Opinions in Multidimensional Topic Spaces. *Phys. Rev. X.* **11**, 011012 (2021,1), https://link.aps.org/doi/10.1103/PhysRevX.11.011012

[11] Cicchini, T., del Pozo, S.M., Tagliazucchi, E. et al. News sharing on Twitter reveals emergent fragmentation of media agenda and persistent polarization. *EPJ Data Sci.* **11**, 48 (2022). https://doi.org/10.1140/epjds/s13688-022-00360-8

Normal intracranial encephalographic activity: oscillations, scale-free behavior and neural network classification

Juan Martín Tenti, Marisa Alejandra Bab, Marcelo Arlego

Keywords: Intracranial, Electroencephalography, Oscillations, Avalanches, Neural Networks

Abstract

In this work, normal intracranial electroencephalographic (iEEG) activity is analyzed in different states of wakefulness and sleep. For this, the database Atlas of iEEG normal activity [1] is used, which compiles iEEG temporal series of patients with refractory epilepsy, during periods free of epileptic seizures, ranging from the waking state and three sleep states: non-REM N2, non-REM N3 and REM.

The aim of the work is to characterize the different states using a variety of tools including frequency bands analysis, the study of scale-free behavior, and classification with neural networks.

The main methods and results are illustrated below, together with the database used.

Database: The Atlas of iEEG normal activity

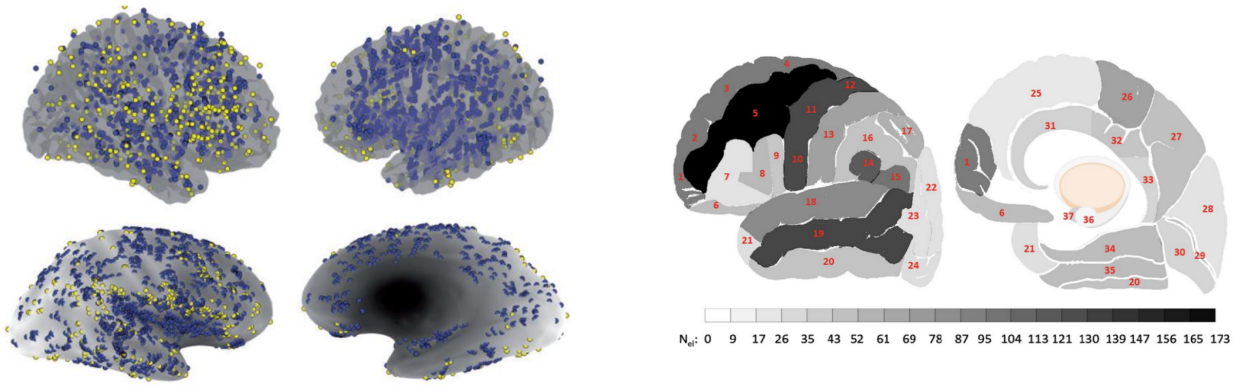


Figure 1. Left: Localization of the 1772 EEG channels with normal physiological activity analyzed for this study. Right: Location of each region and number of electrodes available. The brain areas are labeled in the lateral view (left schematic) and in the medial view (right schematic). See [1] for further details.

Spectral analysis

In the oscillatory analysis (fig. 2) it is observed a very important increase in the spectral power of slow waves during sleep REM, with the N2 phase being more marked. There is a large preponderance of waves in the alpha band during wakefulness, which are not observed during sleep. During non-REM sleep, there is the presence of slow waves in the hippocampus.

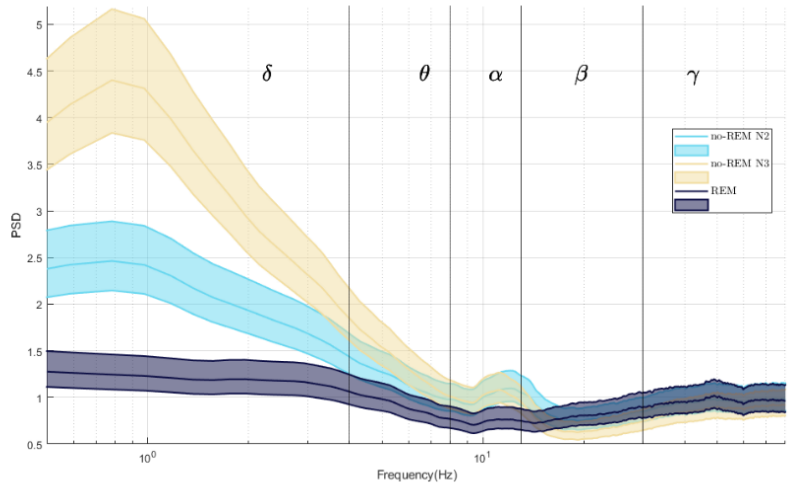


Figure 2: Power spectral density averaged over all subjects for each state of wakefulness and sleep.

Avalanches

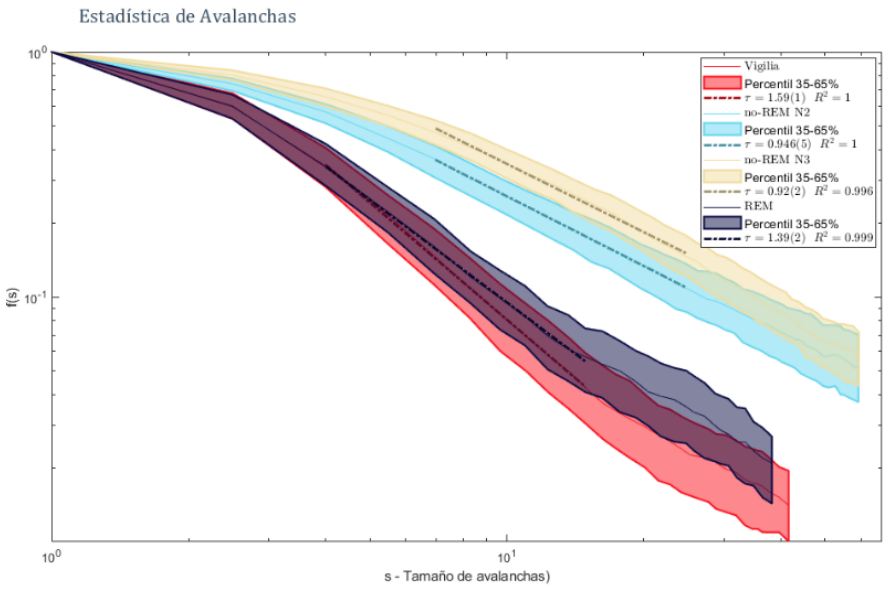


Figure 3: Avalanche distribution for the stages of wakefulness, non-REM N2, non-REM N3 and REM sleep.

The size of an avalanche in iEEG recordings is measured by the number of active electrodes (above a threshold) between instances of inactivity [2]. Figure 3 shows the mean avalanche size distribution on a log-log scale for the different states. The linearity in these curves indicates a trend of scale-free behavior (limited by the accessible scales). As it can be observed, during deep sleep (non-REM) the avalanches are large, covering a large number of electrodes. On the other hand, during wakefulness, there are a large number of small and

local avalanches. Finally, REM sleep has a scale-free behavior very similar to the waking state.

Neural networks

We have used different neural networks to classify sleep-wake states from portions of the iEEG time series or its spectrogram. Figure 4 shows the results of the classification with a convolutional neural network (CNN) [3]. CNN's performance is shown here by means of the Confusion Matrix. The horizontal axis shows the actual states, while the vertical axis shows the predicted states. The best performance here is in predicting the wake state with almost 80% (top left corner). On the other hand, in the last column it can be seen how CNN "confuses" between REM and wakefulness (54% of the REM states are effectively identified as such, but 33% confuse it with wakefulness). This makes sense according to the neurophysiological similarity of both states and is consistent with the results of the other techniques used.

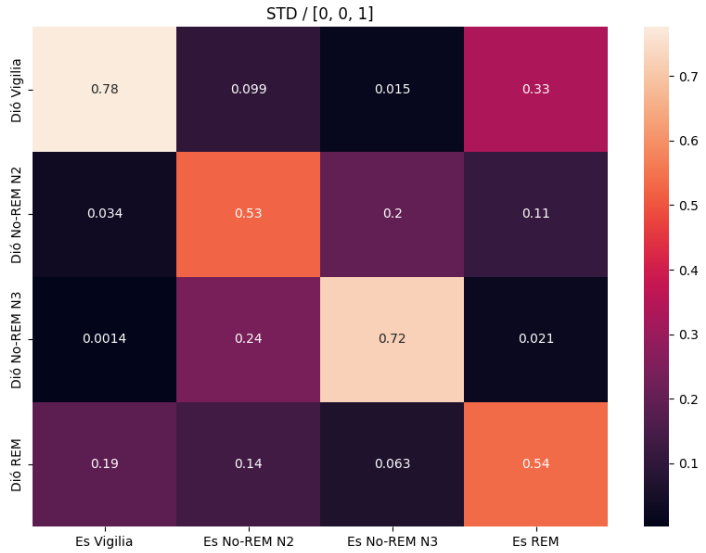


Figure 4. Confusion matrix for state prediction using a convolutional neural network. See text above for description.

Conclusions: Using spectral techniques, scale-free analysis and neural networks we have characterized the different states of wakefulness and sleep in iEEG time series of normal activity. Each technique has its own strengths and limitations, but together they provide a consistent framework that allows classify and identify similarities and differences between states.

References:

[1] von Ellenrieder, N., Gotman, J., Zelmann, R., Rogers, C., Nguyen, D. K., Kahane, P., Dubeau, F., and Frauscher, B. (2020). How the human brain sleeps: direct cortical recordings of normal brain activity. Annals of neurology, 87(2):289–301.

[2] Plenz D. and Niebur, E. (2014). Criticality in Neural Systems. Wiley-VCH Verlag.

[3] Goodfellow I., Bengio Y. and Courville A. Deep Learning. MIT Press, 2016.

On the intricacies of per individual cellular network datasets generation

Anne Josiane Kouam*, Aline Carneiro Viana*, Alain Tchana†,
* *Inria*, France † *Grenoble INP*, France

***Index Terms*—Cellular networks, CDRs, generative models**

I. Introduction

With mobile devices becoming proxies for human presence and activity, datasets collected by mobile operators – i.e., Call Detail Records (CDRs) – are nowadays acknowledged as a common tool to study human behavior in multiple research domains and industries, such as sociology [1], epidemiology [2], transportation [3], and networking [4] (cf. Fig. 1a). CDRs describe time-stamped and geo-referenced event types (e.g., calls, SMS, data) generated by each mobile device interacting with operator networks (cf. Table I). They comprise city-, region-, or country-wide areas and usually cover long time periods (months or years); no other technology provides an equivalent per-device precise scope today.

Yet, the exploitation of real-world CDRs for research faces many limitations (cf. §II). First, *accessibility*: CDRs datasets are not publicly available, imposing strict mobile operators' agreements. Second, *usability*: CDRs are usually available in an aggregated form (i.e., grouped mobility flows and coarse spatiotemporal information), limiting related analyses' preciseness. Third, *privacy*: even anonymized, CDRs describe sensitive information of users' habits, which hardens their shareability [5]. Fourth, *flexibility*: Restricted access to CDRs limits advanced research requiring data richness in terms of population size, duration, or geographical coverage.

This paper introduces the autonomous generation of realistic CDRs to solve the above challenges. In particular, (1) we detail the motivations of such a solution by establishing the scope of such generated traces and describing how it provides new avenues for research advances, and (2) we share our feasibility study of realistic CDRs generation by presenting the related requirements and challenges.

II. Motivation

We first discuss the striking CDRs' research dependencies and the relevance of enabling realistic CDRs' generation.

a) ***CDRs value recognition****:* Generated by the continuous interaction of a urban-wide population with cellular networks, CDRs represent a rich source of knowledge, valuable to many research communities. For a quantitative appreciation, Fig. 1a identifies as many as 14 different research domains leveraging CDRs, among 100 items selected from a 5-year sample set of 1022 publications (gotten from Google Scholar). This clearly shows a great diversity of domains on this sample only ($\sim 10\%$) and considering the 5-year period.

b) ***Limitations in CDRs exploitation****:* Unlike WiFi networks, cellular networks are mobile operators' exclusive property, hardening outside access to collected CDRs. CDRs access is usually granted through NDAs and is often hardly available for most researchers, time demanding, or imprecise due to privacy laws, bringing *accessibility issues*.

Though strongly necessary, privacy compliance asks for CDRs information aggregation, which hardens their usability and limits the exactness of related investigation. Aggregation usually concerns flows, space, time, and event information in CDRs. E.g., the CDRs available at [6] describe aggregated flows of individuals and their number of generated events per intervals of 10 min and square grids of size 235 meters. This points a lack of *information precision* in available CDRs.

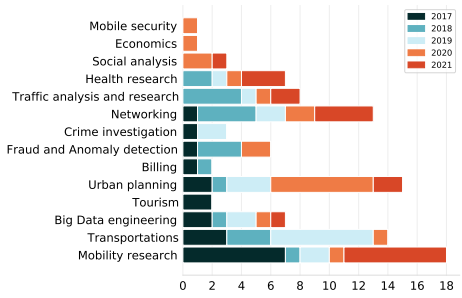
Not surprisingly and justifying the regular CDRs' imprecision issue, personal details of individuals' life habits, inferred from CDRs, calls for privacy-strict exploitation rules and impairs data shareability: e.g., when reconstructed [7] or not [5], majority of individuals' trajectories in CDRs (i.e., higher than 80%) can still be precisely identified, even if anonymized and being sparse in space and coarse in time.

Restricted access to CDRs impacts the *flexibility* of scaling up or adapting CDRs' research results in terms of the population size, the duration, or the covered geographical area, thus limiting advanced research requiring such data richness.

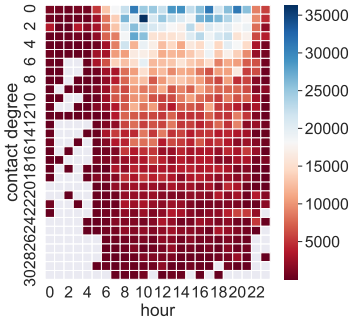
c) ***Conclusion****:* We claim that *an undeniable solution to these limitations is to empower the research community with the flexibility for realistic CDRs' generation that is both adapted to the corresponding research needs and free from the following restrictions: (1) The accessibility to real-world CDRs, (2) The impreciseness of exploiting aggregated CDRs, (3) The impracticality of doing individuals analyses without impeding privacy, and (4) The barrier of not being able to scale up or adapt provided CDRs datasets.* It is worth mentioning the large extent of resulting benefits is likely to profit mobile operators as well as emerging technologies (e.g., 5G/6G), services and applications (e.g., Tactile Internet), or emergencies (e.g., COVID epidemic understanding).

III. CDRs generation requirements

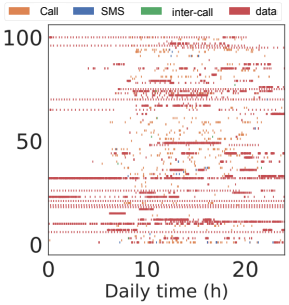
We elaborate on the requirements generated CDRs should meet to ensure broad applicability and reliably reproduce real-world CDRs. First, we present an overview of CDRs gener-



(a) Distribution by domain and year of the most relevant (sorted by Google Scholar) publications using CDRs from 2017 to 2021



(b) Total call duration over a week as a function of contact degree and time



(c) Visualization of temporal sequences of users' events for a real-world CDRs dataset.

Fig. 1

ation; then, we identify five attributes making the generation of realistic CDRs surprisingly challenging and complex.

a) ***Generation overview****:* Mobile traffic generation involves synthesizing timestamped datasets (from x_0 to x_T) only from a given context taken as a parameter, which is much complex traffic prediction, i.e., estimating the next data value at time T based on records from $t = 0$ to $T - 1$. Hence, a generative framework should be expressive enough to fully learn from provided datasets as a training phase, infer the inherent data distributions, and produce new datasets with identical distributions, i.e., realistic. Such frameworks are commonly named *generative models* [8], and while there have been some advances in the literature to efficiently build generative models (e.g., GAN for fake images synthesis), we show in the following the challenges still present for the case of CDRs dataset generation. In particular, we leverage a real-world, fully anonymized CDRs dataset provided by a major network operator in Africa to unveil the intricacies inherent to CDRs, making them difficult to model and generate.

b) ***Modeling inter-features correlations****:* A CDRs dataset comprises both mobility and traffic fields. Mobility fields are essentially user positions (i.e., network cell Ids), while traffic fields are related to network event types, i.e., call, SMS, and data, as described in Table I. CDRs therefore describes three-fold timely behaviors of network users: mobility (*where*), traffic (*what, how*), and social (*whom*) ones. An ideal CDRs generation model should be able to capture the implicit correlations between these features. For instance, Fig. 1b illustrates how social closeness to users' contacts (*social feature*) steadily impacts the hourly duration of calls made (*traffic feature*). Modeling such correlations is unfortunately not a straightforward task, and has never been addressed in the literature. Indeed to the best of our knowledge, state-of-the-art CDRs generation contributions provide the modeling of individual CDRs feature, e.g., [9] for mobility, [10] for data traffic, [11] for social properties.

c) ***Controllability****:* Generated CDRs should be used for a variety of case scenarios (cf. Fig. 1a). The designed generative model should thus allow users to modify the output CDRs by specifying parameters such as the duration, population size, or the mobility area related to a city's urbanization level, layout and infrastructure. Such a controllable generation calls for conditional generative models, rather than just the more common generation approaches based on classical Generative Adversarial Networks (GANs) for instance.

TABLE I: CDRs fields classified into described users' features

	CDR field
General	Phone number
	IMEI
	Timestamp
Traffic	Event-type (call/SMS/data)
	Call duration
	Data session size
Social	Phone number of the correspondent
Mobility	Cell Id

d) ***Modeling arbitrary network topologies****:* CDRs directly reflect the network topology (i.e., cell towers distribution) of its considered mobility area. Hence, there is a strong dependency between operators' CDRs and their network topology, which has to be captured to produce realistic generated CDRs. Such topology however varies with the considered city, requiring the generative model to be able to condition generation on context with arbitrary spatial size. This is a known non-trivial task in machine learning, as popular multilayer perceptron (MLP) or convolutional neural network (CNN) architectures only operate on input with fixed dimensions. More significantly, cellular network topologies are not regular but consist of heterogeneous cells whose shapes vary with the population density of the corresponding covered zones. This makes it impossible to leverage grid tessellations as commonly done in the literature for spatial coverage modeling [12].

e) ***Modeling temporal dynamics****:* Mobile network traffic demonstrates consistent long-term dynamics correlated to regular human activities (peak and off-peak daily hours, weekly working days and weed-ends, yearly vacation and working

months). To some extent, the generated CDRs should faithfully reproduce such dynamics. Unfortunately, this requires first access to long-period-covering real-world data, along with tackling the challenge of learning long-term correlations from acquired datasets. While long short-term memory (LSTM) [13] neural networks are acknowledged as a suitable tool for this latter aspect, the complexity related to multi-various timely dynamics applied to multi-featured CDRs makes the training of such models incredibly thorny.

f) ***Modeling spatiotemporal correlations****:* Mobile traffic datasets include not only spatial or temporal dynamics but also spatiotemporal ones. Specifically, such correlations are induced by human activities in space fluctuations as a function of the time of the day. For instance, in urban life, the office period (9h-17h) presents more traffic events than the after-work one (18h-2h). Such events are concentrated in specific zones corresponding to the working zone of the city. In contrast, the after-work period includes displacement times and night activities, which are not made at specific spots (e.g., people can walk down the streets for their night activity), explaining why events are spread over a broader zone. A good generative model should, therefore, be able to reproduce such spatiotemporal dynamics realistically and regardless of the number of generated users, which is a varying parameter.

g) ***Modeling individuality****:* Last but not the least, reproducing CDRs description per individual demands being able to realistically capture, beyond the global aggregated behavior of the population, the individual behaviors of subscribers in terms of mobility and traffic. While mobility modeling and reproduction is well covered in literature, individuals' cellular traffic reproduction still lacks detailed investigations. In particular, cellular traffic presents a notable heterogeneity that challenges preciseness. As an illustration, Fig. 1c plots the traffic generated during a day by 100 randomly selected users from a real-world CDRs. In the Figure, each line plots a user's sequence of events. We can see a great diversity of users regarding events generation. For example, while some users make predominantly local calls, others make only data; some do not make international calls, and others make it frequently. Similarly, in terms of inter-event time, each user has a singular behavior with events either very timely close together, very far apart, or both. Statistical approaches [10], [14], [15], are limited in reproducing such traffic dynamics as they do not allow per-user modeling but per-user profile (i.e., group of users with similar behavior). This additional challenge refers to multivariate generative modeling, for which current literature is limited to *low-dimensional datasets handling in which each univariate component is independent of the others* [16]. Such techniques are, unfortunately, not adequate for CDRs typically encompassing thousands of network users whose individual timely traffic generations are related by social interactions and to their daily spatiotemporal habits.

IV. Conclusion

Despite the significant value of CDRs datasets, their limited accessibility and usability affect the reproducibility and effectiveness of research in many domains. This paper argues the generation of realistic CDRs is the solution to these limitations. We thus discuss issues to be considered in such solution, which we believe, shed light on the requirements to meet for realistic CDRs generation

References

[1] D. Rhoads, I. Serrano, J. Borge-Holthoefer, and A. Solé-Ribalta, "Measuring and mitigating behavioural segregation using call detail records," *EPJ Data Science*, 2020.

[2] H.-H. Chang, M.-C. Chang, M. Kiang, A. Mahmud, N. Ekapirat, K. Engø-Monsen, P. Sudathip, C. Buckee, and R. Maude, "Low parasite connectivity among three malaria hotspots in thailand," *Scientific Reports*, 2021.

[3] S. Qin, Y. Zuo, Y. Wang, X. Sun, and H. Dong, "Travel trajectories analysis based on call detail record data," in *Chinese Control And Decision Conference*, 2017.

[4] M. Ozturk, A. I. Abubakar, J. P. B. Nadas, R. N. B. Rais, S. Hussain, and M. A. Imran, "Energy optimization in ultra-dense radio access networks via traffic-aware cell switching," *IEEE Transactions on Green Communications and Networking*, 2021.

[5] Y.-A. Montjoye, C. Hidalgo, M. Verleysen, and V. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, 2013.

[6] T. Italia, "Telecommunications - SMS, Call, Internet - MI," 2015. [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV

[7] G. Chen, A. C. Viana, M. Fiore, and C. Sarraute, "Complete Trajectory Reconstruction from Sparse Mobile Phone Data," *EPJ Data Science*, 2019.

[8] H. GM, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, 2020.

[9] M. Zilske and K. Nagel, "Studying the accuracy of demand generation from mobile phone trajectories with synthetic data," *Procedia Computer Science*, 2014.

[10] E. M. R. Oliveira, A. C. Viana, K. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," *Computer Networks*, 2017.

[11] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the structural properties of massive telecom call graphs: Findings and implications," in *ACM CIKM*, 2006.

[12] K. Xu, R. Singh, M. Fiore, M. K. Marina, H. Bilen, M. Usama, H. Benn, and C. Ziemlicki, "Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data," in *IEEE CoNEXT*, 2021.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[14] A. Murtić, M. Maljić, S. L. Gruičić, D. Pintar, and M. Vranić, "Sna-based artificial call detail records generator," in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018.

[15] M. Songailaitė and T. Krilavičius, "Synthetic call detail records generator," *CEUR Workshop proceedings*, 2021.

[16] M. Hofert, A. Prasad, and M. Zhu, "Multivariate time-series modeling with generative neural networks," *Econometrics and Statistics*, 2022.

Outbreak diversity in epidemic waves propagating through distinct geographical scales

Guilherme S. Costa, Wesley Cota and Silvio C. Ferreira

The rapid spreading of COVID-19 mobilized academics from different areas to develop models to understand and predict the behavior of this pathogen. Among the possible approaches, data-driven models, in which a mathematical model is fueled with real data, have proved to be a viable framework to study the epidemic spread of Sars-Cov-2 [1,2]. A central feature of those emerging infectious disease in a pandemic scenario is the spread through geographical scales and the impacts on different locations according to the adopted mitigation protocols. Thus, we investigated a stochastic epidemic model with the metapopulation approach in which patches represent municipalities. Contagion follows a stochastic compartmental model for municipalities; the latter, in turn, interact with each other through recurrent mobility. As a case of study, we consider the epidemic of COVID-19 in Brazil performing data-driven simulations. Properties of the simulated epidemic curves have very broad distributions across different geographical locations and scales, from states, passing through intermediate and immediate regions down to municipality levels. Correlations between delay of the epidemic outbreak and distance from the respective capital cities were predicted to be strong in several states and weak in others, signaling influences of multiple epidemic foci propagating toward the inland cities. The spatio temporal spreading of the pathogen on Brazilian municipalities can be seen in Figure 1 for a weak mitigation scenario. Responses of different regions to the same protocol can vary enormously, implying that the policies of combating the epidemics must be engineered according to the region's specificity but integrated with the overall situation. Real series of reported cases confirm the qualitative scenarios predicted in simulations. Even though we restricted our study to Brazil, the prospects and model can be extended to other geographical organizations with heterogeneous demographic distributions. We acknowledge the funding agencies FAPEMIG, CNPq and CAPES - Finance Code 001

Keywords: Epidemic spreading. Metapopulation. Mobility Network. Data-driven modeling

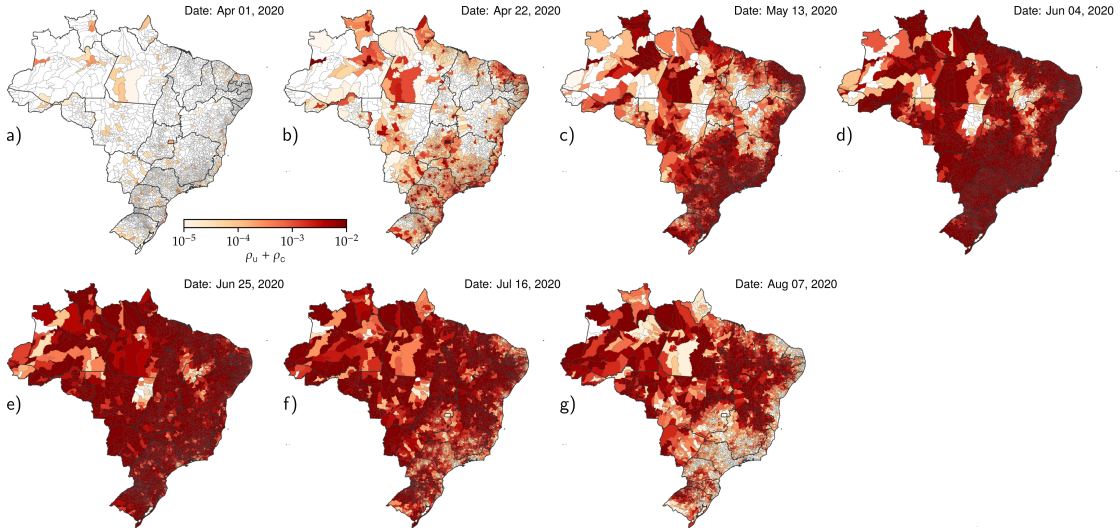


Figure 1: a-g) Color maps presenting the evolution of the prevalence of symptomatic cases for Brazil in a weak mitigation scenario. Dates of the simulations are shown in the upper right corner of each frame. The darker colors represent higher prevalences in a logarithm scale.

References

1. L. Danon, E. Brooks-Pollock, M. Bailey, and M. Keeling, "A spatial model of covid-19 transmission in england and wales: early spread, peak timing and the impact of seasonality," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 376, no. 1829, p. 20200272, 2021.

2. A. Arenas, W. Cota, J. Gómez-Gardeñes, S. Gómez, C. Granell, J. T. Matamalas, D. Soriano-Paños, and B. Steinegger, "Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions," Phys. Rev. X, vol. 10, p. 041055, Dec 2020.

Quantifying Biobank Impact

Rodrigo Dorantes-Gilardi[1] and John Michael Gaziano[4,5] Albert-László Barabási[1,2,3]

[1] Network Science Institute, Northeastern University, Boston, USA
barabasi@gmail.com
[2] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[3] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[4] Department of Internal Medicine, Harvard Medical School, Boston, MA, USA
[5] Massachusetts Area Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston, MA 02130, USA

1 Introduction

Biobanks are repositories conceived for epidemiological research of biological samples and the associated data from a large group of individuals. During the last few decades, the use of biobanks has increased to the extent of being a fundamental component of biomedical research and has the potential to significantly improve future healthcare [1]. In 2009, Time magazine added biobanks as one of the ten ideas that were changing the world, and their creation has been promoted by research centers and gained international support from funding agencies and the research community as a whole [2]. In the last decade, several grants have been obtained for the purpose of creating, maintaining, or relying on biobank resources [3].

Biobanks vary widely in terms of purpose, scope, governance, and type of data. The team composition of a biobank can also differ in terms of gender composition, popularity of its lead, and size of the team. It is not surprising that their impact in science is dissimilar, and only a handful of well-known biobanks receive most of the necessary attention to be highly impactful in terms of publications.

2 Results

The characteristics of high-impact biological cohorts and the mechanisms of recognition given to their creators remain elusive, however, as quantitative measures applied to the universe of biobanks are hard to find. Here, we use a data mining approach to identify a list of more than one thousand biobanks together with their introductory paper to the academic community (Figure 1A). We use the corpus of articles to track their footprint on research under different angles, making it to our knowledge, a first case of a large-scale quantitative study of biobank academic impact. We show that biobank impact is unbalanced across the world by first describing the origin and usage of biobanks. Most of them are originated and used in the global north, and regions like south-east Asia and Africa have a usage rate smaller than their production rate.

In order to study the different methods biobanks' teams receive recognition from researchers, we measure the extent that biobank data-access results in collaboration with external researchers. On average, 56% of the citations to biobanks come through collaborations with third party authors (Figure 1B). In the case where data access is more restricted, we observe that a larger share of publications is co-authored with external researchers, and then discuss that stringent data sharing policies may be set in place to avoid lack of recognition. We use the UK Biobank case to observe that roughly half of the papers identifying the biobank as a data source in the article or abstract do not cite any of the resource articles written by the biobank team leading to reduced measurable impact (Figure 1C).

Finally, we implement two logistic regression models to predict the academic impact of a biobank based first on its inherent characteristics (Figure 2A, 2C), like cohort size and data openness, and another one based on the academic rank and gender composition of the team of the biobank (Figure 2B, 2D). We find that biobanks with genetic data for general purpose research (as opposed to disease specific) tend to be more impactful; unexpectedly, we observe that restrictive biobanks tend to have more citations as well. Biobank impact increases as the share of female team members increases, proving that highly diverse teams are good for biobank formation. The popularity of the lead author at the time the biobank was created is a major influence of its success.

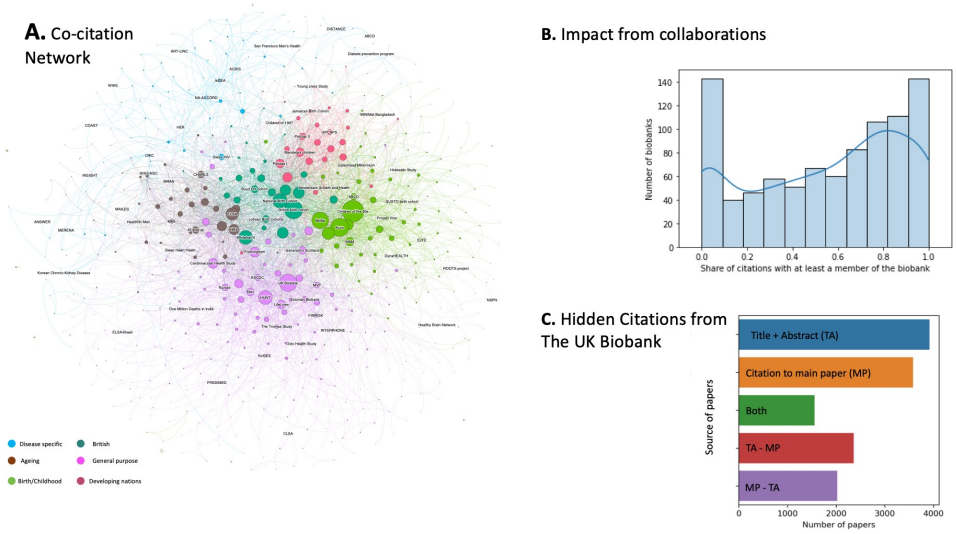
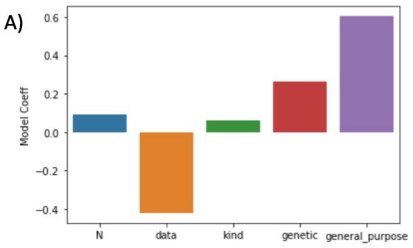
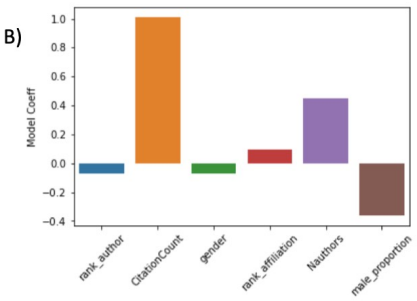


Fig. 1. Biobank Impact A. Biobanks represented by their co-citation network where nodes are biobanks that are connected if they are cited by the same paper. Different communities are given a color and represent a cluster of biobanks. B. Distribution of the share of citations from collaborations (at least one member of the biobank team is co-author of the citing article). C. Hidden citations of the UK Biobank are defined as the number of articles using the data of the biobank without citing any of the biobank's articles





C) Inherent properties model

	A	B	C
N	6	3	5
Data	0	1	0
Kind	0	1	1
Genetic	1	0	0
General purpose	1	0	1
P(Success)	0.72	0.36	0.66

D) Lead author model

	A	B	C
Rank	1k	10k	10k
cites	100k	10k	100k
gender	male	female	male
Rank affiliation	1k	10k	10k
Nauthors	8	16	8
Male proportion	50%	20%	75%
Pr(Success)	0.69	0.58	0.65

Fig. 2. Coefficients of the logistic regression models to predict biobank impact. A. Model is based on the characteristics of the biobank, namely, sample size (N), data openness, kind, genetic data available (genetic), and whether it is general purpose or disease specific. B. The model is based on the characteristics of the biobank's team, these include the rank of the lead scientist of the team, its popularity (based on citations received at the time the biobank was created), gender, and the rank of their affiliation, the team size of the biobank, and the male proportion of the team. Three examples of biobanks are shown in C) ad D) to represent different biobanks and their probabilities of being successful (highly cited).

References

1. Shilo, Smadar, Hagai Rossman, and Eran Segal. ”Axes of a revolution: challenges and promises of big data in healthcare.” Nature medicine 26.1 (2020): 29-38
2. Caulfield, Timothy, et al. ”A review of the key issues associated with the commercialization of biobanks.” Journal of Law and the Biosciences 1.1 (2014): 94-110
3. Kinkorová, Judita, and Ondřej Topolčan. ”Biobanks in Horizon 2020: sustainability and attractive perspectives.” Epma Journal 9.4 (2018): 345-353

Quantifying individual uncertainty in decision-making: Unrelated preferences for degree programs reduce students' first-year retention in higher education.

Cristian Candia [1,2], J. Davyt-Colo [3], M. Guevara [4], J. Pulgar [5], T. Yaikin [3], C. Monge [6], F. Pinheiro [7], C. Rodriguez-Sickert [3]

[1]Data Science Institute, Facultad de Ingeniería, Universidad del Desarrollo, Las Condes, 7610658, Chile.

[2]Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL 60208, USA

[3]Centro de Investigación en Complejidad Social, Facultad de Gobierno, Universidad del Desarrollo, Chile.

[4]Department of Computer Science, Universidad de Playa Ancha, Chile.

[5]Departamento de Física, Universidad del Bío Bío, Concepción 4051381, Chile

[6]Feedback comunicaciones, Vitacura, Chile.

[7]Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisboa, Portugal

Uncertainty describes the quality of our information concerning risk for a given decision. For those decisions affecting our lives in the long-term, we usually invest time exploring and sampling information to reduce such uncertainty. Higher education is a milestone in people's lives. Therefore, sampling information regarding degree programs is pivotal to reap long-term educational outcomes. We propose a new framework that quantifies individual uncertainty based on a network structure of degree programs, the Higher Education Space (HES). We build the HES using data on 1.6 million applicants' preferences in Chile between 2005 and 2019 (top panel, Figure 1). Then, we quantify individual uncertainty by computing the relatedness of degree program preferences (Rk, Eq. 1) using:

$$R_k = \frac{1}{N(N-1)} \sum_{i \neq j \in P_k} d_{ij}, \qquad (1)$$

where N is the number of degree programs in the application (P_k) of individual k, and d_{ij} is the network distance between degree programs i and j in application P_k. To evaluate the impact of uncertainty in individual decision-making, we test whether the relatedness between applied degree programs impacts first-year retention (the continued enrollment in the same degree program). Note that not-retention is arguably a costly outcome for both individuals and institutions. We find that, on average, students who select related programs have a 73% first-year retention probability, which becomes 40% when applying to unrelated degree programs (bottom panel, Fig. 1). This effect is steeper for high-score applicants (yellow curve, bottom panel, Fig. 1). Indeed, the retention rate equalizes between low and high-score applicants when applying to distant programs. The results are robust to all the available socio-economic control variables such as family income, enrolled institution, province of origin, province of enrolment, and family size, among many others. Besides, we replicate our results using data from Portugal and a regression discontinuity design quasi-experiment for causal inference. Finally, we build prediction models for classifying students in a potential academic risk in an early stage (the beginning of their academic year) with an 80% accuracy. Thus, we provide a network-science framework for quantifying uncertainty in decision-making processes that impact long-term outcomes. In this case, our framework can help prevent academic dropouts by identifying students at risk early and then to focalize institutional accompanying programs such as tutoring, vocational interventions, or remedial classes.

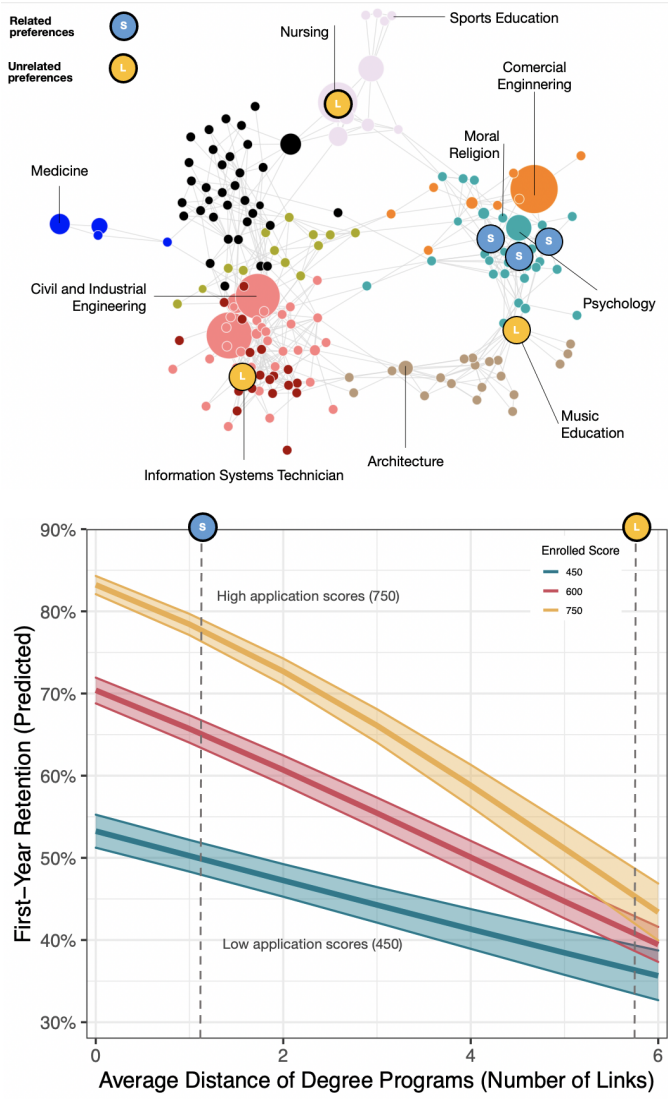


Fig. 1: The top panel shows Higher Education Space (HES) and degree program relatedness. The bottom panel exposes the impact of average distance degree programs on predicted retention probability.

Random multi-player games in complex networks

Keywords: Evolutionary games, complex networks.

Authors: Natalia Kontorovsky, Juan Pablo Pinasco, Federico Vázquez.

Evolutionary game theory provides a framework to study the behavior of large populations where individuals playing different strategies (or having different biological traits) interact through some game, and they can replicate according to their payoffs.

Evolutionary games with pairwise interactions were extensively studied [1, 2]. Local interactions with various opponents had also been considered, where in each round the same fixed number of players are randomly selected from the population to play against each other. However, there are many situations in which the number of players can vary over time and even between rounds. It also happens that the optimal strategies in a two players game could not be optimal in a three players game, thus interactions between multiple players can not be reduced to pairwise interactions. Therefore, it is interesting to model and study the case in which the game can be played by a different number of players in each round when the strategy must be selected previously, without knowing a priori the exact number of players involved.

The concept of evolutionary stable strategy (ESS) is a central when studying time evolution, because it satisfies the additional stronger condition of stability, which implies that if an ESS is reached, then the proportions of players playing the different strategies do not change over time

In this work [3] we formalize and generalize the definition of evolutionary stable strategy (ESS) to be able to include a scenario in which the game can be played by a different number of players in each round. Even though a similar problem was analyzed previously in terms of two types of players, incumbents (original population) and mutants (invaders)[4], only two combinations of them were considered. Here we show that all combinations must be considered, and a hierarchy of payoffs is needed in order to characterize an ESS when the number of players in each interaction is a random variable.

In order to explore these questions, we study the simplest non-trivial case of the duel-truel game. As usual, in a duel two players aim to eliminate each other, while in a truel three players are involved. Each player can use one of two possible strategies, that we call *perfect and mediocre* strategies. When a player uses the strategy perfect, then it annihilates its competitors with probability 1.0, while a player using strategy mediocre kill its opponents with probability 0.5. The paradox is that, when a truel game is played, a perfect player is not necessarily the winner of the game, even having the highest killing probability.

This surprising result was already present in the early literature on truels [5]. As a consequence, in a duel game the ESS corresponds to the entire population playing strategy perfect, whereas in a truel game the ESS corresponds to all players using strategy mediocre.

This led us to consider what the ESS would be in a scenario where the number of players is a random variable, the so-called Poisson games, where at each iteration step of the dynamics a duel is played with probability $p \in (0, 1)$, and a truel is played with the complementary probability $1 - p$.

We also introduce an agent-based model in which players interact in a complex network by copying the strategies of their neighbors, with a dynamics that in mean-field evolves following the replicator equation. Let us observe that the replicator dynamics is usually defined in terms of new individuals entering the population, by selecting a pure strategy with a probability proportional to the payoff given the current mix of agents. Our approach has an independent interest, and has the advantage that can be used in networks with a fixed number of nodes or agents, bypassing the issue of how to add new nodes as agents replicate.

We perform extensive Monte Carlo (MC) simulations of the model in different types of networks, and develop an analytical approach based on a pair approximation (PA) that allows to obtain approximate equations for the evolution of the fraction of perfect agents in the network. This approach enable us to investigate if the transitions between ESS in pure and mixed strategies found within the Nash equilibrium theory are also observed in complex networks, and to identify how the networks' topology affects the existence of mixed equilibria.

In Fig. 1 we can see that the coexistence of perfect and mediocre players predicted by the mean-field approach when interactions are all-to-all (dashed line) is also present when agents interact in complex topologies (solid lines), and have a good agreement with MC simulations (symbols). Figure 2 shows the phase diagram in the p–μ space, where μ is the mean degree of the network. We observe that the coexistence phase shrinks as μ decreases, but it does not seem to vanish completely even for small values of μ. As a consequence, a given unstable mix of the two types of players for some value of p, can turn into stable when the mean number of neighbors of a player is increased beyond a threshold. This result implies that the network of interactions affects the stability of the system by inducing a stable coexistence when its connectivity increases.

References

[1] J. P. Pinasco, M. Rodriguez Cartabia, and N. Saintier, "Evolutionary game theory in mixed strategies: From microscopic interactions to kinetic equations," Kin. Relat. Models 14, 115–148 (2021).

[2] J. M. Smith, "Evolutionary game theory," Physica D 22, 43–49 (1986).

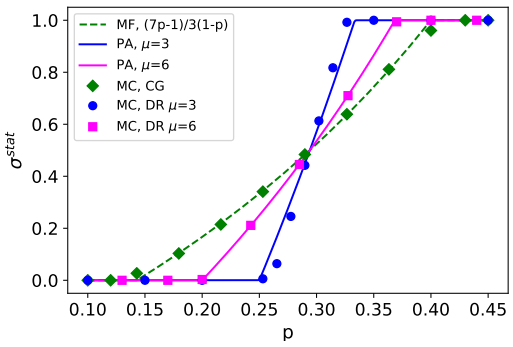


Figure 1: Stationary fraction of perfect agents σ^{stat} vs duel probability p, for the values of the mean degree μ indicated in the legend. The dashed line corresponds to the stable solution σ_{CG} on a Complex Graph (CG), while solid lines represent the solution from the PA equations. Symbols correspond to the average value of σ at the stationary state obtained from MC simulations on a CG of size $N = 10^3$ (diamonds), and DRRGs of size $N = 10^4$ and degrees $\mu = 6$ (squares) and $\mu = 3$ (circles).

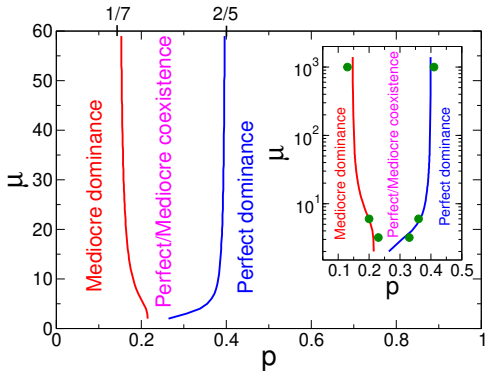


Figure 2: Phase diagram on the p–μ space showing the transition lines between the coexistence and dominance phases, obtained from the PA equations for an ER network.

[3] Kontorovsky, N. L., Pinasco, J. P., Vazquez, F. (2022). Random multiplayer games. Chaos: An Interdisciplinary Journal of Nonlinear Science, 32(3), 033128.

[4] H. Tembine, E. Altman, R. El-Azouzi, and Y. Hayel, "Evolutionary games with random number of interacting players applied to access control," in 2008 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops (IEEE, 2008) pp. 344– 351.

[5] P. Amengual and R. Toral, "Truels, or survival of the weakest," Computing in Science Engineering 8, 88–95 (2006).

Reconstructing social sensitivity from evolution of content volume in Twitter

Sebastián Pinto,[1,2,*] Marcos A Trevisan,[1,2] and Pablo Balenzuela[1,2]

[1] *Departamento de Física, FCEN, Universidad de Buenos Aires. Pabellón 1, Ciudad Universitaria, 1428EGA, Buenos Aires, Argentina.*

[2] *Instituto de Física de Buenos Aires, CONICET. Ciudad Universitaria, 1428EGA, Buenos Aires, Argentina.*

KEYWORDS

Public interest; media coverage; opinion dynamics.

EXTENDED ABSTRACT

In this work, we set up a simple mathematical model for the dynamics of public interest in terms of media coverage and social interactions. We test the model on a series of events related to violence in the US during 2020, using the volume of tweets and retweets as a proxy of public interest, and the volume of news as a proxy of media coverage. The model succesfully fits the data and allows inferring a measure of social sensibility that correlates with human mobility data. These findings suggest the basic ingredients and mechanisms that regulate social responses capable of ignite social mobilizations.

Our approach is grounded in the Granovetter model [1], originally proposed to explain the emergence of riots. In this model, agents adopt a binary state s which we interpret as interested ($s = 1$) or non-interested ($s = 0$) in the event. The dynamics of the system is described in terms of the *public interest*, the fraction $p = \sum_i^N s_i/N$, where N is the size of the system. Each agent is characterized by a threshold τ_i, which is the fraction of interested agents needed to induce interest on the agent. Thresholds are random variables whose cumulative distribution $S(p) = P(\tau < p)$ is interpreted here as *social engagement*, given that it represents the fraction of agents that become active due to their threshold lies below p. Assuming that thresholds are normally distributed $\tau \sim N(\mu, \sigma)$, we have:

$$S(p|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{p} e^{-\frac{(\tau-\mu)^2}{2\sigma^2}} d\tau. \qquad (1)$$

When μ is low, small groups can trigger the interest to the rest of the system. On the contrary, high values of μ would require a bigger fraction of interested people to induce interest to rest of the population. We therefore identify the quantity $1-\mu$ as the *social sensitivity* of the population. In his original model, Granovetter described the dynamics of the public interest p regardless of the influence of the media. To include this, we propose a modified model that analytically reads:

$$\frac{1}{\gamma}\frac{dp}{dt} = -p + e\,C(t) + (1-e)\,S(p|\mu(t),\sigma). \qquad (2)$$

Equation 2 allows us to reconstruct the social variables from data. By integration this equation using the volume of twitted news as a proxy for the coverage $C(t)$, we seek for the functions $S(t)$ that minimize the difference between the resulting public interest and the volume of tweets and retweets. In panel (A) of figure 1 shows two of the analyzed cases: one related to the murder of George Floyd in May 2020 and, the other, the attack against Jacob Blake in August 2020 (other events was analyzed in [2]). Panel (B) of figure 1 shows the best fitting curves for the public interest and the reconstructed social engagement and social sensitivity for both cases. As can be seen in this figure, the two social variables are of a different nature. In fact, while the engagement $S(t)$ is a threshold-based variable whose dynamics can be expected to be fast, $1-\mu(t)$ represents the slower, more gradual build-up of social sensitivity across the whole population. Accordingly, we find that this variable changes appreciably over periods of ~ 15 days which is, as expected, longer than the typical time scales of the media coverage and public interest.

Panel (B) of figure 1 also shows periods of time of increasing social sensitivity, which leads to a sudden increase of the social engagement, when a macroscopic fraction of agents becomes interested in the events. We expect that this increament impacts beyond the digital environment. Therefore, we investigate the emergence of measurable collective activity associated to an increase in social sensitivity by collecting mobility measures across the US territory. In panel (C) of figure 1 we show attendance to recreation places, groceries, pharmacies and public transport stations in the period of time when the events took place (Minneapolis, Minnesota for the case of George Floyd, and Kenosha, Wisconsin for Jacob Blake). We find different degrees of correlation between the social sensitivity and mobility patterns for the most populous events using a lag of 3 days.

Taken together, these results suggest that our low-dimensional approximation of the Granovetter model

* spinto@df.uba.ar

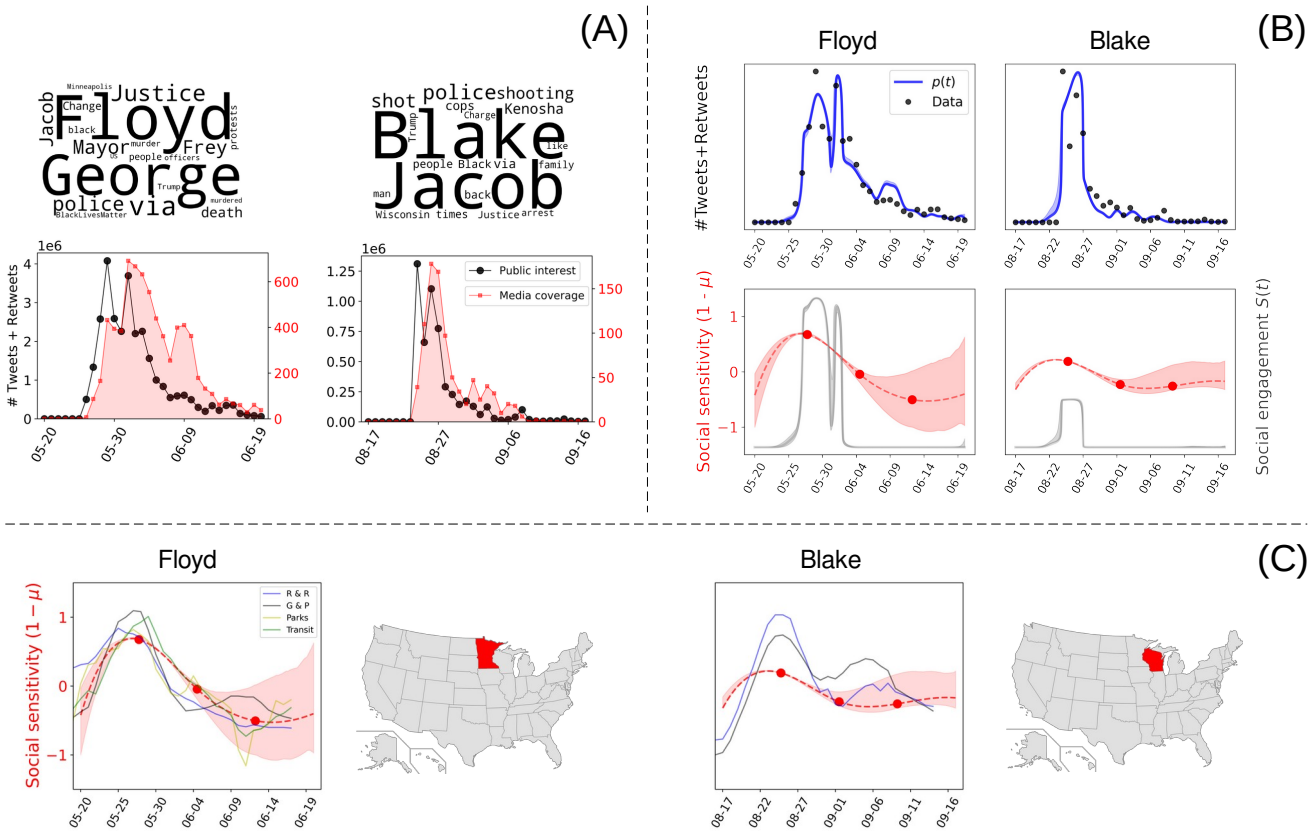


FIG. 1. **Panel(A)** Time traces of the volume of tweets and retweets (black circles, same data as panel (b)) and media accounts tweets (filled area). **Panel(B)** In the top panel points correspond to public interest (tweets and retweets) along with the best fitting curves $p(t)$ (blue) obtained with the model of equations 2 and 1; bottom panel shows in red lines social sensitivity $1-\mu(t)$, while in grey lines the normalized social engagement $S(t) = S(p|\mu(t), \sigma)$. **Panel (C)** Social sensitivity (red) and standardized mobility observables of the corresponding county. R & R: retail and recreation; G & P: groceries and pharmacies; Parks: public parks; Transit: transit in public transport stations (Parks and Transit not shown for Blake due to lack of data). All mobility measures were shifted -3 days and inverted for visualization purposes.

captures the basic ingredients that regulate social responses of very different magnitudes, which are indeed capable of ignite social mobilizations. The model implements the hypothesis that agents become involved from media exposure and also from the presence of a critical mass of interested agents in the system, which leads to characterize the social sensitivity of the population. Further details and the analisis of other events can be found in [2].

[1] M. Granovetter, Threshold models of collective behavior, American journal of sociology **83**, 1420 (1978).
[2] S. Pinto, M. Trevisan, and P. Balenzuela, Reconstructing public engagement from social media content volume, arXiv preprint arXiv:2112.11644 (2021).

Rumor-telling activity in polarized opinion networks

Hugo P. Maia, Silvio C. Ferreira and Marcelo L. Martins
Universidade Federal de Viçosa, Brazil

Keywords: Complex Networks, Sociophysics, Rumor propagation, Opinion polarization.
Preprint avaliable at arxiv [1].

In the past, rumors have ignited revolutions, undermined the trust in political parties, or threatened the stability of human societies. With the constant development of online social networks, rumors propagation and fake-news dissemination are becoming ever potentially dangerous for harmonious living in a society of differing opinions. Rumors are frequently affected by whatever alignment the population has with relation to it.

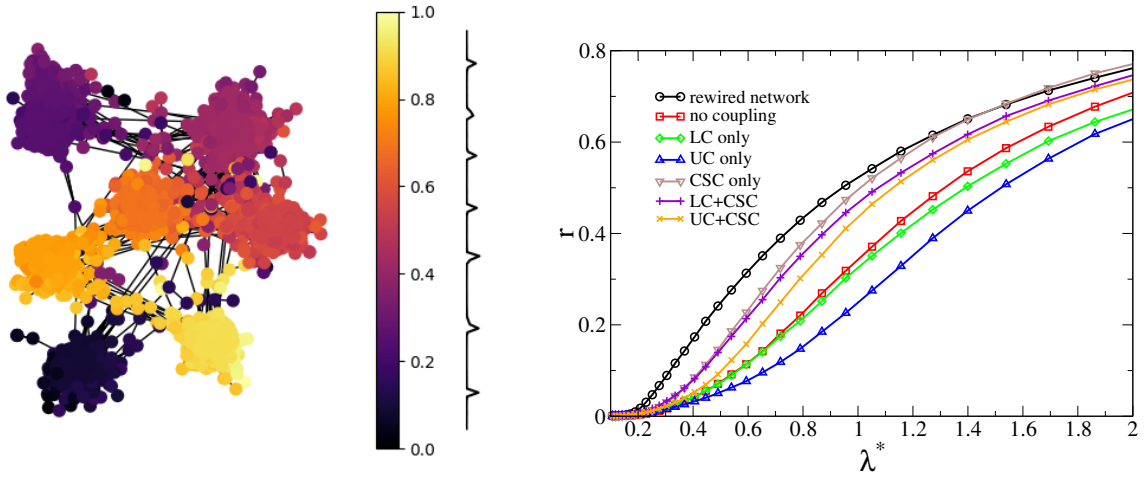
Several theoretical and empirical studies have been devoted to understanding rumor-spreading dynamics [2, 3]. Recent empirical works have observed that the structure of online communication networks frequently exhibits echo chambers, in which beliefs are reinforced due to repeated interactions with individuals sharing the same points of view [4, 5]. Moreover, these communities form a new topological structure of the communication network as weakly interconnected modules. So, besides the heterogeneous degree distribution, there is an additional level of heterogeneity associated with community sizes in modular networks.

We investigate rumor spreading models on complex networks generated by an adaptive opinion formation processes [6] that leads to loosely connected modular networks forming echo chambers (as shown in Fig. 1(a)). Here, rumors are coupled with the opinion of the interacting agents according to different rules that alter the individual's spreading rate λ_i and each link's stifling rate α_{ij}, leading to a modified rumor model. The model for the dynamics of spreaders (I), ignorants (S) and stiflers (R) can be summarized as follows:

$$I_i + S_j \xrightarrow{\lambda_i} I_i + I_j \qquad I_i + R_j \xrightarrow{\alpha_{ij}} R_i + R_j \qquad I_i + I_j \xrightarrow{\alpha_{ij}} R_i + I_j$$

A linear coupling (LC) between rumors and opinions was investigated, following from the assumption that if a rumor is ideologically aligned to an individual's opinion, he will be more prone to disseminate it. An unimodal coupling (UC) was also investigated, in which not only individuals aligned to the rumor are more likely to spread it, but also individuals of opposite alignment who spread the rumor as a criticism. In addition, a controversy-seeking coupling (CSC) where contrasting opinions hampers lost of interest on an issue was also proposed and investigated.

We show that the highly modular structure of opinion polarized networks strongly impairs rumor spreading. However, the introduction of couplings between agent's opinions and their spreading/stifling rates has a striking effect on rumor-telling. In Fig. 1(b), the final fraction of stiflers is shown for different combinations of couplings to compare the general rumor spreading capability as a function of the overall individual's average spreading rate λ^*. Indeed, depending on the nature of the couplings, information propagation can be either further inhibited or enhanced up to the level observed in unpolarized networks, thus suppressing the modularity bottleneck. Information percolation and permeability are also studied for the analysis of how rumors originating from an extreme alignment can affect those of opposite alignment. The time to reach the absorbing state and a variability analysis for the final density of stiflers were also studied for different combinations of couplings and compared to unpolarized networks, revealing that controversy-seeking behavior not only is capable of overcoming the bottleneck hurdles, but also drastically extends the rumor's lifespan.



(a) (b)

Figure 1: (a) Typical network obtained with the opinion formation model of Ref. [6]. Colors represent individual's opinions. The opinion distribution is presented besides the color bar; the latter indicate the opinion scale in range $[0, 1]$. (b) Final fraction of stiflers r as a function of λ^* for different types of opinion couplings. Rewired networks are used as a control case, the opinion distribution is kept but links are reshuffled while preserving the degree distribution in order to eliminate the network modular structure. Acronyms: LC - linear coupling; UC - unimodal coupling; CSC - controversy-seeking coupling.

References

[1] H. P. Maia, S. C. Ferreira, and M. L. Martins, "Controversy-seeking fuels rumor-telling activity in polarized opinion networks," 2022.

[2] D. J. Daley and D. G. Kendall, "Epidemics and rumours," *Nature*, vol. 204, p. 1118, 12 1964.

[3] D. Maki, M. P, M. Thompson, P. Hall, and T. M, *Mathematical Models and Applications: With Emphasis on the Social, Life, and Management Sciences*. Prentice-Hall, 1973.

[4] N. Gaumont, M. Panahi, and D. Chavalarias, "Reconstruction of the socio-semantic dynamics of political activist twitter networks-method and application to the 2017 french presidential election," *PLOS ONE*, vol. 13, pp. 1–38, 09 2018.

[5] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, p. 35, 2019.

[6] H. Maia, S. Ferreira, and M. Martins, "Adaptive network approach for emergence of societal bubbles," *Physica A: Statistical Mechanics and its Applications*, vol. 572, p. 125588, 06 2021.

Scale-Free Network without a Power-Law Degree Distribution

Xiangyi Meng[1] and Bin Zhou[2]

[1]*Network Science Institute and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA*

[2]*School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China*

Keywords: Scale-free network, Degree–degree distance, Preferential attachment, Fitness

Originating from the ideas of renormalization group in statistical field theories [1], the study of the *scale-free* property in network theories has undergone an impressive development over the past twenty years [2, 3]. It is noteworthy that the term "scale-free" [2] was first dubbed to denote the unique property that a power-law distribution of, e.g., degree k, $f(k) \propto k^{-\alpha}$ is invariant (free) under the continuous scale transformation $k \to k + \epsilon k$. However, since a pure power-law degree distribution (DD) would not be normalizable in the domain $k \in [0, \infty)$, the DD cannot truly be power law but requires an ultraviolet (UV) cutoff, in terms of either a $k_{\min}$ (the minimum degree each node can have) or some other nontrivial corrections around small k. The DD regains its scale invariance only asymptotically in the infrared (IR) limit $k \to \infty$.

Therefore, a finite-size network can only be approximately "scale-free" [4]. This has put great difficulty in how we can test if real-world finite-size complex networks are "scale-free" in *abundance*, a remarkable claim that, albeit supported by many empirical observations [5], remains controversial in recent literature [6]. Without prior knowledge of the UV cutoff, it is unknown how large a typical degree k must be, that we should consider as already entering the power-law regime where rigorous statistical analysis can be employed. An oversimplified solution to this is to statistically test the full domain of k, ignoring possible UV cutoffs, resulting in that less than one third of real-world networks have a statistically significant power-law DD [7]. Yet, when another metric is investigated, namely the degree–degree distance η [7], defined by

$$\eta = \exp|\ln k_i - \ln k_j| \tag{1}$$

for every link $i \leftrightarrow j$ connecting two nodes i and j, it turns out that real-world networks almost universally have a statistically significant power-law degree–degree distance distribution (DDDD) [7]. Much interest has since been drawn to the characteristics of η, bringing its prevalence to broader network science topics [8], especially network closeness [9] and network assortativity [10].

Despite the potentials, the finding of power-law DDDD raises a question: *is there any theoretical relevance between the power laws (if any) of DD and DDDD?* It appears that asymptotically, a power law of DDDD $g(\eta) \propto \eta^{-\beta}$ is nothing but a delegate of the power law of DD of the same network, given that an equality $\beta = \alpha - 1$ [7] has been derived for both the Barabási–Albert (BA) model [2] and a special power-law-distributed fitness model [7]. There seems no reason to investigate the power law of DDDD for its own theoretical purpose.

Nevertheless, here we show that the power law of DDDD is *more than* a delegate. Our main result is that the set of networks with an asymptotic power-law DD is a *proper subset* of those with an asymptotic power-law DDDD. This immediately indicates that there are networks whose DD is not power law, but DDDD is, differing not only in statistical significance but also in their asymptotic limits. This also implies that our current understanding of the scale-free property of networks in terms of only the power law of DD is incomplete. Indeed, given the broader scope of power-law DDDD, we propose that it would be more appropraite and general to denote "scale-free networks" as having a power-law distribution for *any* of its metrics, not the degree only. Such, the scale-free property need not manifest in all metrics, thus better identified and distinguished not by apparent power laws but by the underlying network mechanisms, such as *preferential attachment* [2] or *quenched fitness* [11]. In particular, we will show that either of these two fundamental mechanisms can generate networks as concrete examples for our purpose—representing scale-free networks without a power-law DD.

Results.—We (abusively) phrase our main result as

$$\mathcal{D}^2|_{\text{power-law}} \subset \mathcal{D}^4|_{\text{power-law}} \tag{2}$$

and derive it in two steps, claiming its (i) *inclusion* and, more interestingly, its (ii) *strict inequality* as follows:

(i) $\mathcal{D}^2|_{\text{power-law}} \subseteq \mathcal{D}^4|_{\text{power-law}}$, i.e., *every network with a power-law DD also has a power-law DDDD.*

(ii) $\left|\mathcal{D}^2|_{\text{power-law}}\right| < \left|\mathcal{D}^4|_{\text{power-law}}\right|$, i.e. *there are networks that do not have a power-law DD but exhibit a power-law DDDD.* We will consider two network models of general interest:

Preferential attachment of internal links only.—As our first model, this "no-growth" model differs from the BA model [2] in that the number of nodes N of the network is fixed as a constant, and only internal links are added to the initially empty network during its evolution. At each time step t, two nodes i and j are randomly and independently chosen, resulting in a link drawn between

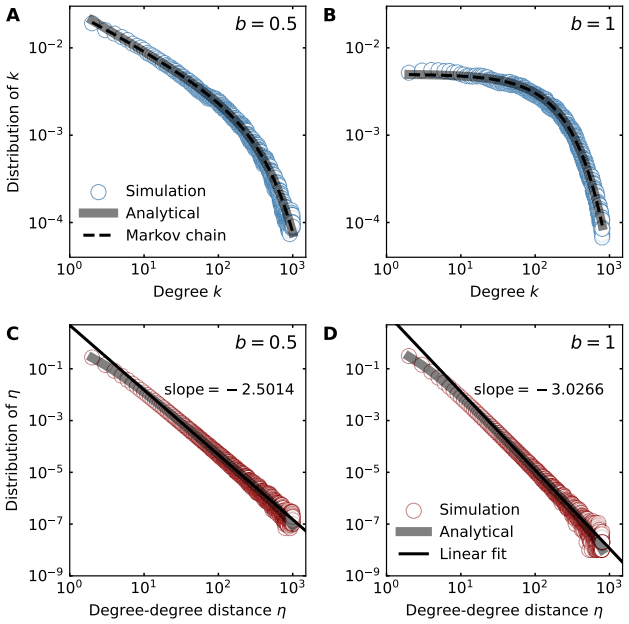


FIG. 1: Distributions of (a-b) k and (c-d) η of the "no-growth" model, generated by preferential attachment of internal links only (with attachment probability $\propto k +$ a small constant b). Simulation results (circle): average of 10^2 runs on $N = 10^4$ nodes and $T = 10^6$ links.

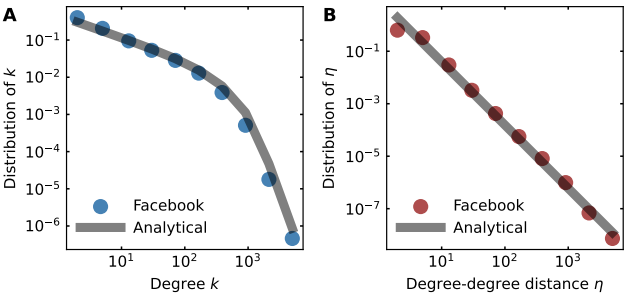


FIG. 2: Distributions of (a) k and (b) η of the Facebook network, fitted by analytical results (solid line) of the "no-growth" model with $b = 0.4092$.

i and j. The probability of choosing such two nodes is $\propto (k_i + b)(k_j + b)$, i.e., preferentially proportional to each node's current degree k plus a small constant b. After T time steps, the network acquires T links.

We derive the analytical results for both DD and DDDD of the no-growth model. In contrast to DD, we find that the DDDD exhibits a strong power law in the small b regime. The analytical result matches the simulation result [Fig. 1(c-d)], with the power-law exponent of DDDD equal to $2 + b$ as expected.

It is worth noting that as $b \to 0$, we rediscover the classic scaling $\sim \eta^{-2}$ for the DDDD of the BA model [7]. This is evidence that the scale-free property of the BA model indeed originates from the preferential attachment

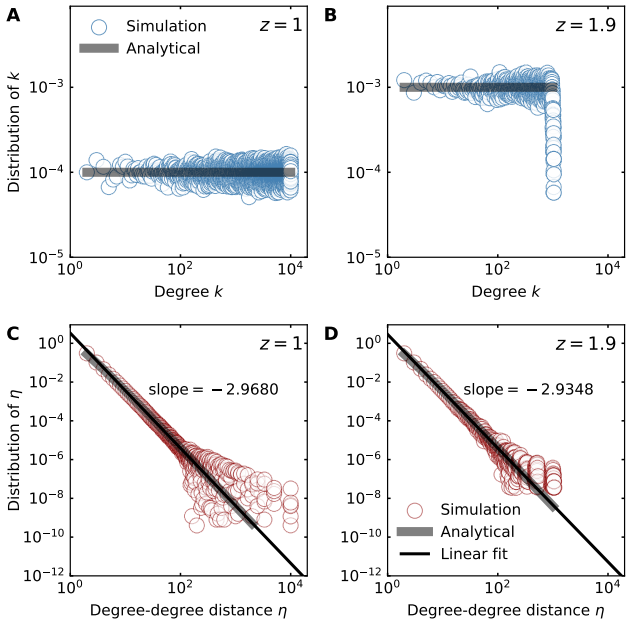


FIG. 3: Distributions of (a-b) k and (c-d) η of the "strong-coupling" model, a uniform-distributed fitness model (fitness $\omega \in [0, \omega_{\max} = 1]$) where two nodes are linked if their fitness sum is larger than a threshold z. Simulation results (circle): average of 10^2 runs on $N = 10^4$ nodes.

mechanism but *not* from network growth, a discrimination that cannot be revealed by comparing DD only [12].

Facebook is the world's largest social networking platform. Figure 2 shows the DD and DDDD of Facebook. We find that the DD of Facebook is not a power-law distribution, but DDDD is. Notably, both DD and DDDD are also in good agreement with the analytical results of the no-growth model, suggesting that the model is more than a toy model but of strong practical significance as well.

Fitness with threshold.—The second model of interest is defined by assigning a random fitness ω that follows a fitness distribution $\rho(\omega)$ to each node of the network [13]. For every two nodes i and j, let a link be drawn with probability $\sigma(\omega_i, \omega_j)$ that depends only on the fitness of the nodes, not their degrees [11]. By choosing $\sigma(\omega_i, \omega_j) = \theta(\omega_i + \omega_j - z)$, where $\theta(x)$ is the Heaviside step function, a strong coupling of fitness is introduced, such that a link will be deterministically drawn if and only if the sum of the fitnesses of the two nodes is larger than a threshold z. In particular, here we consider a uniform distribution $\rho(\omega)$ for $\omega \in [0, \omega_{\max}]$ and assume that $\omega_{\max} \leq z \leq 2\omega_{\max}$. We find that the annealed average DD is simply given by $f(k) \sim N^{-1}$ [Fig. 3(a-b)] following the calculation in Ref. [14]. For DDDD, we derive $g(\eta) \sim \eta^{-3}$ for large η, a strong power law that is independent of z [Fig. 3(c-d)].

Discussion.—We also observe that Eq. (2) establishes

an inclusive order between the power laws of DD and DDDD. This raises the question of whether we can find another more inclusive metric beyond DDDD. It would be interesting if a hierarchy between all such metrics could be established, especially for scale-free networks, that offers new insights on distinguishing the origins of scale-free properties.

[1] C. Itzykson and J.-M. Drouffe, *Statistical Field Theory: Volume 1, From Brownian Motion to Renormalization and Lattice Gauge Theory*, 1st ed. (Cambridge University Press, New York, 1989); *Statistical Field Theory: Volume 2, Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems*, 1st ed. (Cambridge University Press, New York, 1989).

[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[3] L. A. Adamic and B. A. Huberman, Science **287**, 2115 (2000); P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000); P. L. Krapivsky, G. J. Rodgers, and S. Redner, **86**, 5401 (2001); C. Song, S. Havlin, and H. A. Makse, Nature **433**, 392 (2005); T. S. Evans and J. Saramäki, Phys. Rev. E **72**, 026138 (2005); S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008); F. Radicchi, J. J. Ramasco, A. Barrat, and S. Fortunato, Phys. Rev. Lett. **101**, 148701 (2008); M. Szell, R. Lambiotte, and S. Thurner, Proc. Natl. Acad. Sci. **107**, 13636 (2010); R. Lambiotte, M. Rosvall, and I. Scholtes, Nat. Phys. **15**, 313 (2019); T. Nesti, F. Sloothaak, and B. Zwart, Phys. Rev. Lett. **125**, 058301 (2020).

[4] M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar, and G. Caldarelli, Proc. Natl. Acad. Sci. **118**, e2013825118 (2021).

[5] A.-L. Barabási and E. Bonabeau, Sci. Am. **288**, 60 (2003); K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, Proc. Natl. Acad. Sci. **99**, 12583 (2002); R. Pastor-Satorras, E. Smith, and R. V. Solé, J. Theor. Biol **222**, 199 (2003); R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000); K.-I. Goh, B. Kahng, and D. Kim, **87**, 278701 (2001).

[6] A. D. Broido and A. Clauset, Nat. Commun. **10**, 1017 (2019); P. Holme, **10**, 1016 (2019); R. E. Langendorf and M. G. Burgess, Sci. Rep. **11**, 20501 (2021); M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar, and G. Caldarelli, Proc. Natl. Acad. Sci. **118**, e2013825118 (2021).

[7] B. Zhou, X. Meng, and H. E. Stanley, Proc. Natl. Acad. Sci. **117**, 14812 (2020).

[8] B. Wang, J. Zhu, and D. Wei, Mod. Phys. Lett. B **35**, 2150331 (2021); A. J. Maren, Entropy **23**, 319 (2021).

[9] T. S. Evans and B. Chen, Commun. Phys. **5**, 1 (2022).

[10] A. Farzam, A. Samal, and J. Jost, Sci. Rep. **10**, 1 (2020).

[11] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).

[12] A.-L. Barabási, *Network Science*, 1st ed. (Cambridge University Press, Boston, 2016).

[13] G. Bianconi and A.-L. Barabási, EPL **54**, 436 (2001).

[14] V. D. P. Servedio, G. Caldarelli, and P. Buttà, Phys. Rev. E **70**, 056126 (2004).

Score-driven Generalized Fitness Model for Sparse and Weighted Temporal Networks

Domenico Di Gangi[1], Giacomo Bormetti[2], Fabrizio Lillo[2,3]

[1] ISTI-CNR, via G. Moruzzi 1, 56124 Pisa
[2]Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126, Pisa, Italy
[3]Department of Mathematics, University of Bologna,
Piazza di Porta San Donato 5, 40126 Bologna, Italy

Keywords— Temporal Weighted Networks, Fitness Models, Score Driven Models, Inter-Bank Market

Introduction In the last two decades, networks, or graphs, have attracted an enormous amount of attention as an effective way of describing pairwise relations in complex systems. The ever increasing abundance, and variety, of graph data has motivated a great deal of applications of statistical models to graphs [14]. More recently, the availability of time varying networks' data has stimulated the development of temporal network models [11, 15]. While the vast majority of this literature focuses on *binary graphs*, i.e. graphs that are defined solely by a set of nodes and a set of links between pairs of nodes, often we can associate a weight to each link. In such cases the data is better described by a weighted, or valued, network, that can be described with a positive, real valued matrix $Y_{ij} \in \mathcal{R}^+$. Y_{ij} is the value of the link between node i and node j, and $Y_{ij} = 0$ if the link is not present.

One important well known fact is that real world valued networks are very often found to be sparse, i.e. their adjacency matrices have an abundance of zero entries. That is the case, for example, of inter-bank networks [2], a class of weighted temporal networks of paramount importance [1], that are known to be extremely relevant to financial stability [10], and have motivated the application and development of a number of statistical models for networks [3, 13].

Contribution Highlights Our main contribution is a model for sparse weighted dynamical networks, that also accommodates for the dependency of the network dynamics on external variables, and its application to weighted temporal network data, describing overnight exposures in the European inter-bank market. Our work contributes to the extremely scarce literature on dynamical models for sparse weighted networks by extending the very well known fitness model for static binary networks. We consider a zero augmented generalized linear model to handle the weights and a state of the art econometric approach to describe time varying parameters. This results in a flexible model that allows us to decouple the probability of a link to exist from its expected weight, and to explore the influence of external regressors on the network's dynamics. We then exploit such flexibility to investigate how the relevance of EONIA rates on the eMid interbank market changed over time.

This work has been recently published in [7] and is related to [6] where we generalize the family of Exponential Random Graph Models (ERGM) for binary networksto its score driven version (SD-ERGM).

Methods' Overview We start from the well known *fitness model* [4, 8], also known as *beta model* or *configuration model*. That is a model for binary networks where, in the case of directed networks, each node i is assigned two parameters, the in fitness $\overleftarrow{\theta}_i$ and the out fitness $\overrightarrow{\theta}_i$, that capture the tendency of each node to form incoming and outgoing connections. We consider a slightly generalized binary fitness model, that allows for the dependency on external regressors X_{ij}, where the probability of a link to exist is described by

$$p_{ij} = \frac{1}{1 + e^{-(\overleftarrow{\theta}_i + \overrightarrow{\theta}_j + X_{ij}\beta_{bin})}}. \tag{1}$$

Alongside the standard, *binary fitness* parameters, we associate to each node i two new parameters $\overleftarrow{\varphi}_i$, $\overrightarrow{\varphi}_i$, that we call *weighted fitness*. They describe the propensity of a node to have more or less heavy weights in incoming and outgoing links respectively, and use them to model the weighted adjacency matrix Y_{ij} with the

Model	T_{train}	MSE Log.	MAD Log.	AUC
Localized Tobit	100	2.351	1.067	0.714
Localized Tobit	200	2.267	1.043	0.795
SD Generalized Fitness	100	0.859	0.726	0.896

Table 1: Results of the link and weights prediction exercise on e-MID data. We compare the Localized Tobit model of [9] with our proposed score-driven generalized fitness model.

following Zero Augmented distribution

$$P\left(Y_{ij}=y\right)=\begin{cases}\left(1-p_{ij}\right) & for \quad y=0\\ \\ p_{ij}\; g_{ij}\left(y\right) & for \quad y>0\end{cases}, \tag{2}$$

where $g\left(y\right)$ is the density for a positive continuous random variable, for example the gamma distribution, defined such that

$$E\left[Y_{ij}|y>0\right]=e^{\left(\overleftarrow{\varphi}_i+\overrightarrow{\varphi}_j+X_{ij}\beta_w\right)}.$$

We then extend this model to the dynamical context by allowing the fitness, both binary and weighted, and the β parameters to change over time, following the Score Driven approach introduced by [5] and [12]. Given a sequence of observed weighted adjacency matrices $\left\{\mathbf{Y}^{(t)}\right\}_{t=1}^{T}$, and denoting by $f^{(t)}$ a, K dimensional, vector containing all time varying parameters $\overleftarrow{\theta}$, $\overrightarrow{\theta}$, β_{bin}, $\overleftarrow{\varphi}$, $\overrightarrow{\varphi}$, β_w, we assume that $f^{(t)}$ follows a score-driven recursive update rule

$$f^{(t+1)}=w+\boldsymbol{\beta}f^{(t)}+\boldsymbol{\alpha}\boldsymbol{S^{(t)}}\frac{\partial\log P\left(\mathbf{Y}^{(t)}|f^{(t)}\right)}{\partial f^{(t)\prime}}, \tag{3}$$

where w, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are static parameters, estimated via maximum likelihood, w being a K dimensional vector and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ $K\times K$ matrices. $\boldsymbol{S^{(t)}}$ is a $K\times K$ scaling matrix. We run extensive numerical simulations to make sure that this update rule defines an effective way to *filter* the time varying parameters, also when their temporal evolution is governed by a different Data Generating Process.

As an empirical application, we study a portion of the European interbank market (e-Mid) where nodes are banks and a directed and weighted link represents a loan between two banks. Our dataset contains the list of all credit transactions each day from June 6, 2009, to February 27, 2015 and, using a weekly aggregation, this amounts to a sequence of T=298 networks. We perform a link and weight forecasting analysis and we compare our results with a state of the art econometric model for temporal networks [9] which uses a Localized Tobit regression on past network metrics and a local likelihood estimation. Table 1 shows the comparison between the two models for what concerns the link prediction (using the Area Under the Curve as metric) and the weight prediction (using the Mean Squared Error and Mean Absolute Error of the logarithm of the loan size). It is evident that our model outperforms the competitor along both dimensions, even when the latter is allowed to be trained on a longer dataset.

Finally, we consider a model variation which consider the EONIA interest rate as an additional driver of the network dynamics finding that an increase in this reference rate increases the probability of a link but decreases this weight. This important interpretation is not obtained by using a fitness model without the score driven dynamics.

References

[1] Franklin Allen and Ana Babus. "Networks in Finance". In: *The network challenge: strategy, profit, and risk in an interlinked world*. Ed. by Paul R JKleindorfer, Yoram Jerry R Wind, and Robert E Gunther. Social Science Research Network, 2011. Chap. 21, pp. 367–379.

[2] Kartik Anand et al. "The missing links: A global study on uncovering financial network structures from partial data". In: *Journal of Financial Stability* 35 (2018), pp. 107–119.

[3] Leonardo Bargigli et al. "The multiplex structure of interbank networks". In: *Quantitative Finance* 15.4 (2015), pp. 673–691.

[4] G. Caldarelli et al. "Scale-Free Networks from Varying Vertex Intrinsic Fitness". In: *Physical Review Letters* 89 (25 2002), p. 258702.

[5] Drew Creal, Siem Jan Koopman, and André Lucas. "Generalized autoregressive score models with applications". In: *Journal of Applied Econometrics* 28.5 (2013), pp. 777–795.

[6] Domenico Di Gangi, Giacomo Bormetti, and Fabrizio Lillo. "Score-Driven Exponential Random Graphs: A New Class of Time-Varying Parameter Models for Dynamical Networks". In: *arXiv:1905.10806* (2019).

[7] Domenico Di Gangi, Giacomo Bormetti, and Fabrizio. Lillo. "Score-driven Generalized Fitness Model for Sparse and Weighted Temporal Networks". In: *Information Sciences (in press)* (2022).

[8] Diego Garlaschelli and Maria I. Loffredo. "Maximum likelihood: Extracting unbiased information from complex networks". In: *Physical Review E* 78 (1 2008), p. 015101.

[9] Liudas Giraitis et al. "Estimating the dynamics and persistence of financial networks, with an application to the Sterling money market". In: *Journal of Applied Econometrics* 31.1 (2016), pp. 58–84.

[10] Andrew G Haldane and Robert M May. "Systemic risk in banking ecosystems". In: *Nature* 469.7330 (2011), pp. 351–355.

[11] Steve Hanneke, Wenjie Fu, Eric P Xing, et al. "Discrete temporal models of social networks". In: *Electronic Journal of Statistics* 4 (2010), pp. 585–605.

[12] Andrew C Harvey. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Vol. 52. Cambridge University Press, 2013.

[13] Piero Mazzarisi et al. "A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market". In: *European Journal of Operational Research* 281.1 (2020), pp. 50–65.

[14] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[15] Daniel K Sewell and Yuguo Chen. "Latent space models for dynamic networks". In: *Journal of the American Statistical Association* 110.512 (2015), pp. 1646–1657.

Self-induced consensus formation among Reddit users on the GameStop short squeeze

Anna Mancini,[1,2] Antonio Desiderio,[1,2] Riccardo Di Clemente,[3,4] and Giulio Cimini[1,2]

[1] *Physics Department and INFN, Università degli Studi di Roma Tor Vergata, 00133, Rome, Italy*

[2] *Centro Ricerche Enrico Fermi, 00184, Rome, Italy*

[3] *Department of Computer Science, University of Exeter, United Kingdom*

[4] *The Alan Turing Institute, London NW12DB United Kingdom*

Keywords: Online Social Networks, Opinion Dynamics, Consensus Formation

The GameStop short squeeze of January 2021, primarily orchestrated by amateur investors on the Reddit r/wallstreetbets (WSB) community, represents an unprecedented example of a collective coordination action on online social media, with tangible impact on the stock market. A theoretical knowledge of the microscopic dynamics that led to this event is still lacking, but empirical evidence suggests that a fundamental understanding of the GME case study requires an endogenous self-reinforcing mechanism able to trigger consensus formation. In this work [1] we characterize the structure and time evolution of WSB conversations, identifying early signs of collective action that can be associated to an increasing level of user commitment towards the short squeeze operation.

The first variable we looked at is the occurrence of stock tickers in the text of posts and comments. As Figure 1A shows, peaks in the Z-scores for "GME" occurrences correspond to all major events of the GameStop saga, and they become more frequent in time, signaling a growth in interest towards GME. It is also noteworthy that peaks of GME occurrences and trading volume of the stock mostly coincide, pointing to a strong relation between the two variables. Another variable of interest is the *sentiment* of comments, that has been analyzed using VADER (*Valence Aware Dictionary and sEntiment Reasoner*)[2]. Figure 1B shows the average sentiment of all daily posts/comments that mention GME. We see that as early as the beginning of December the trend starts to grow significantly (with respect to the baseline which refers to the whole conversation). We can interpret this empirical evidence as a growing commitment towards the GME operation, representing an early sign of consensus formation in the community.

In light of these results, we worked out a model in which user engagement can influence collective behavior and foster the emergence of consensus. We build on the *voter model*, following an approach formally similar to [3]. There are two opinions in the population, ± 1, which in our context can be

associated with participation or not to the short squeeze operation. At each time step a user is selected at random; with probability $1-\lambda$ she copies the state of a random neighbor j, whereas, with probability λ she follows a global field given by a random variable $e(t) = \pm 1$, depending on a control parameter $c \geq 1$ associated with the level of user engagement. Studying the analytical mean-field solution of the model, in Figure 1C we can see how the system exhibits a classic second order phase transition at the critical value $c = e^2$: below this threshold no opinion prevails, whereas, above the threshold the dynamics quickly reaches a stable equilibrium point $|m^*| \neq 0$ that becomes closer to full consensus as c grows. We proceeded on studying how the model behaves on user-user interaction networks extracted from WSB conversation data, by building a user network for each month, from October 2020 to January 2021. Since the success of the short squeeze required a large number of investors who bought and held GME shares, we introduce an extensive order parameter: $M(t) = m(t)\, N_0 e^{qm(t)}$, which grows in time to mimic the many new users who joined the community in correspondence of the squeeze. This extensive magnetization exhibits a sharp transition, as shown in Figure 1D for the January user network, properly describing a sudden and large-scale formation of consensus qualitatively similar to the abrupt growth of GME price, which ultimately represents the best proxy for the success of the short squeeze.

[1] Mancini, A., Desiderio, A., Di Clemente, R. & Cimini, G. *Self-induced consensus of Reddit users to characterise the GameStop short squeeze*. Scientific Reports 12:13780 (2022).

[2] Hutto, C. & Gilbert, E. *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. In Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, 216–225 (2014).

[3] De Marzo, G., Zaccaria, A. & Castellano, C. *Emergence of polarization in a voter model with personalized information*. Phys. Rev. Res. 2, 043117 (2020).

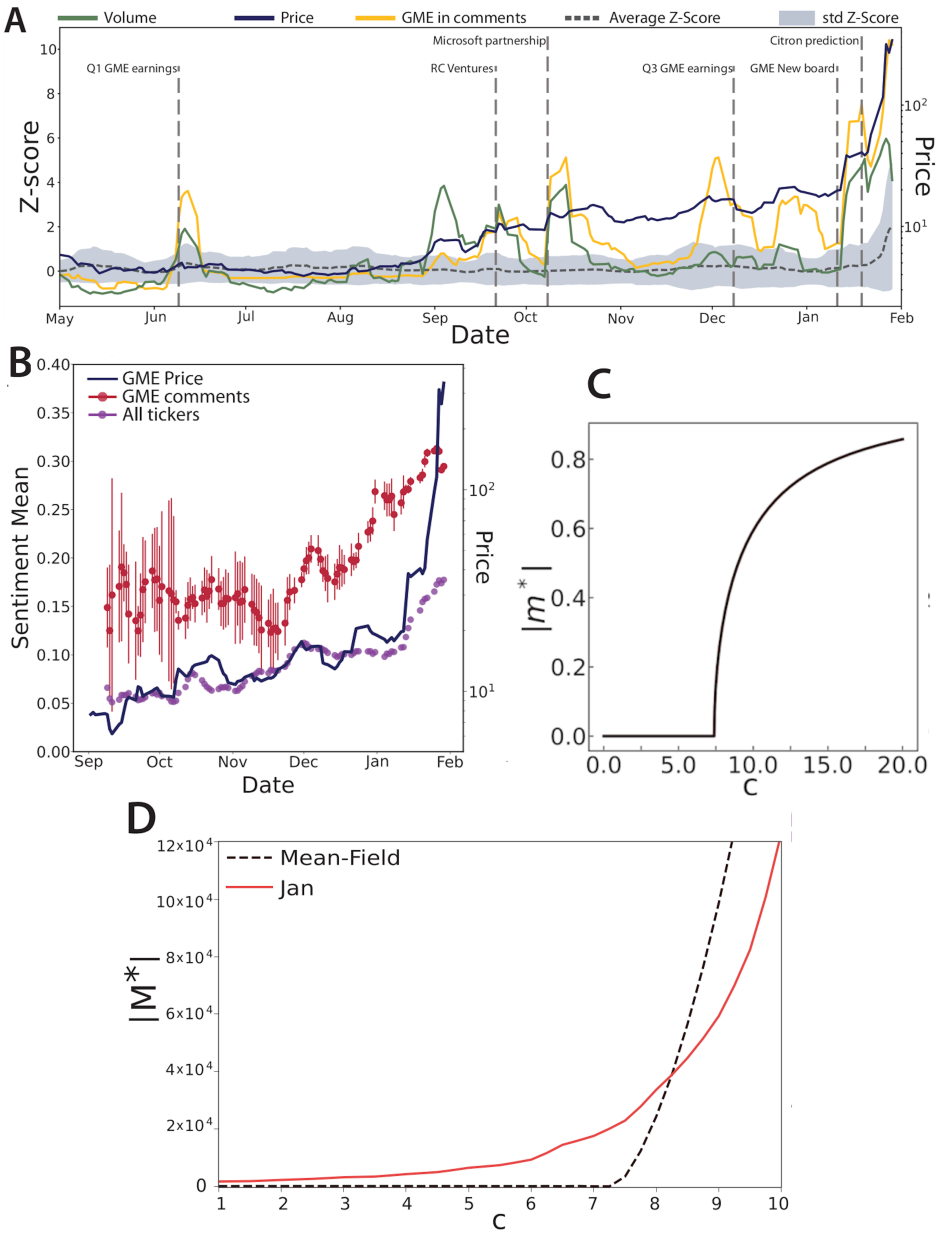


FIG. 1. A) Z-score for the occurrences of "GME" in users comments, compared to the mean Z-score for the occurrences of all stock tickers, and Z-score of the volume of transactions of GME shares. B) Average sentiment of comments containing "GME", with respect to the same quantity computed over all comments and GME closing price. C) Equilibrium points of the magnetization $|m^*|$ as a function of the control parameter c (the level of user engagement), according to mean field approximation. D) Phase transition of the extensive order parameter in the model and for empirical user-user network in Januray.

SPATIOTEMPORAL ANALYSIS OF EARTHQUAKE OCCURRENCE IN SMALL-WORLD-LIKE OFC MODEL SYNTHETIC DATA AND ACTUAL EARTHQUAKES

Douglas Ferreira[1,2], Jennifer Ribeiro[2,3], Paulo Oliveira Jr[4], André Pimenta[2], Renato Freitas[2], Rafael Dutra[2], Andrés Papa[3,4], and José Mendes[1]

[1] *Department of Physics, University of Aveiro, Aveiro, Portugal*
[2] *LISComp – Laboratory, Federal Institute of Rio de Janeiro, Paracambi, RJ, Brazil*
[3] *Department of Geophysics, National Observatory, Rio de Janeiro, RJ, Brazil*
[4] *Department of Physics, Rio de Janeiro State University, Rio de Janeiro, RJ, Brazil*

Keywords: Earthquakes; Nonextensive statistical mechanics; Small-world networks; Olami-Feder-Christensen model

The study of earthquakes is of great importance for several knowledge areas, given their high destructive power. When studying seismic events a handy tool is the use of computer simulations, which enable the creation of synthetic data that mimics the real ones. Among the existing models to generate earthquakes, the one created in 1992 by Olami, Feder and Christensen [1], known as the OFC model, stands out for its simplicity and capacity of reproducing several statistical properties of actual earthquakes.

The standard OFC model can be represented by a bidimensional square L x L lattice with $N=L^2$ blocks interconnected to its first neighbors by springs. Each block is connected through a spring to a single rigid driven plate and by friction to another rigid fixed plate on which they stay. This blocks arrangement represents a regular topology of the lattice. Due to the relative motion between the plates, that is imposed by the model, all the blocks will be subjected to an elastic force which tends to put them in motion, and to other frictional forces, opposite to the first. When the resulting force in one of the blocks is greater than the maximum static friction force, the block slides and relaxes to a position of zero force, and a fraction α, of its tension is equally redistributed between its nearest neighbors. It produces a rearrangement of forces in its first neighbors, which can cause other slippages and the emergence of a chain reaction. The first block to move is the earthquake's epicenter, and the magnitude s of this earthquake is measured by the number of blocks that skidded.

When an earthquake finishes, the continuous relative movement between the plates will cause a force accumulation in all blocks. In this way, the slip of a block will occur after some time, and a new earthquake process begins. In addition, the model assumes that the time interval between two earthquakes is considerably longer than the duration of an earthquake itself. However, despite the good results provided by this model, the authors in [2] showed that a better agreement with real data is obtained when the lattice's topology in the OFC model is not regular, but rather follows the rule of topology construction based on the Watts–Strogatz mechanism [3] to generate small-world networks [4]. In this way, starting from the regular two-dimensional topology [Fig. 1(a)], the rule of construction consists of taking each edge of the lattice (the spring that connects the blocks) and randomly reconnecting it to other blocks with probability of rewiring p, keeping fixed the original number of connections of each block. This mechanism generates a *small-world-like* lattice for the OFC model, in which the resultant lattice has a topology between the regular and the random ones [Fig. 1(b)].

The use of a small-world-like topology for the OFC model makes the system more realistic since it allows more effective long-range interactions between the blocks, agreeing with several previous works which indicate spatial and temporal long-range interactions between earthquakes. In the present work, we analyzed the spatiotemporal features of the synthetic earthquake data generated with this modified OFC model through distributions of distance ($\triangle r$), time interval ($\triangle t$) and "propagation" of the seismic activity (given by a quantity similar to velocity, $v = \triangle r/\triangle t$) between successive earthquakes, and compared the results with those obtained for actual worldwide earthquakes. In all cases we varied the lower size threshold, s_{th}, of the simulated earthquakes and the

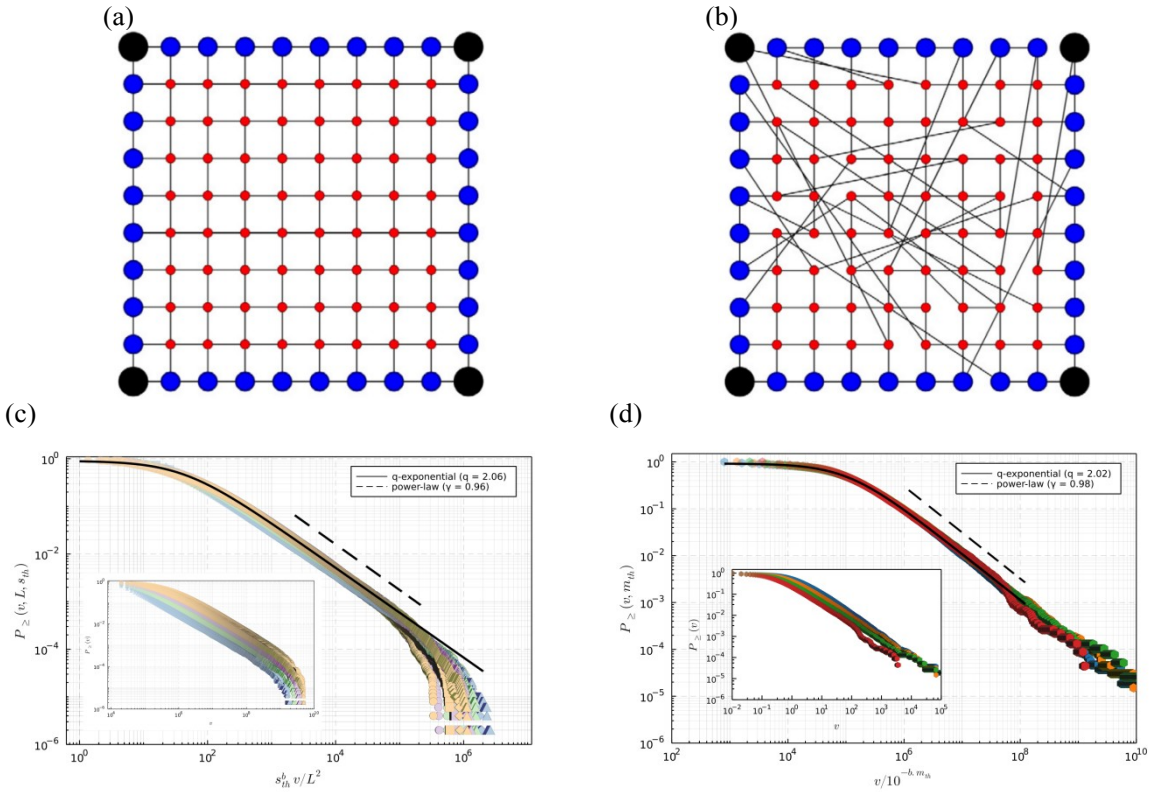


Fig. 1 – (a) An example of a 10 x 10 OFC model lattice with a regular topology and (b) with a small-world topology with p = 0.05. In (b), the presence of long-range correlations allows connections between distant blocks. (c) Data collapse in the cumulative probability distribution of velocities for consecutive synthetic earthquakes from the modified OFC model. The lattice sizes considered are L = 200 (light blue), 300 (light green), 400 (light purple), and 500 (light yellow). The lower size thresholds are, s_{th} = 25 (circle), 50 (square), 75 (diamond), and 100 (triangle). The solid black line is a *q-exponential* function with q = 2.06 and β_v = 2.262 x 10^{-2}. The dashed black line shows the power-law behavior. (d) Data collapse of the distribution of velocities for successive worldwide shallow earthquakes. The velocities are in km/s. The magnitude thresholds, m_{th}, considered are 4.5 (blue), 4.7 (orange), 4.9 (green), and 5.1 (red). The solid black line is a *q-exponential* with q = 2.02 and β_v = 9.087 x 10^{-6}. The dashed black line shows the power-law regime. *Insets:* same distributions, but without data collapse.

lattice sizes, L, of the OFC model, thus, the probability distributions are functions of L and s_{th}.

We noted that all distributions have a good agreement with the non-traditional function *q-exponential*, that belongs to the nonextensive statistical mechanics theory [5] and is defined by

$$e_q(x) = \begin{cases} [1+(1-q)x]^{1/(1-q)} & \text{if} \quad [1+(1-q)x] \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

This result is in concordance with previous studies for earthquakes from certain regions of the world, such as California, Japan, Iran and Greece [6-8]. Furthermore, we found that all distributions can be made invariant with respect to the values of s_{th} and L when applying scaling laws similar to the ones found for earthquakes from specific locations of the planet [9-11]. This result indicates the presence of criticality in the earthquakes and spatiotemporal self-similarity. Fig. 1(c) shows the data collapse and the *q-exponential* fit for the cumulative probability distribution of velocities, of the form $P_{\geq}(v) = e_q(-\beta_v v)$, where β_v is a scale constant, for successive synthetic earthquakes from the modified OFC model.

The agreement with *q-exponential* functions and the existence of scaling relationships that make the distributions invariant with respect to the magnitude thresholds, m_{th}, considered were also found for the distributions of distance, inter-event time and velocity for the worldwide earthquakes. We highlight that the global actual data used in this study was divided in shallow earthquakes (those with a depth up to 70 km) and in deep earthquakes (the ones localized at depths greater than 70 km), since they are mechanically different from each other [12,13]. For exemplification, Fig. 1(d) presents the result obtained for the distribution of velocities considering consecutive shallow seismic events.

Our results reinforces the conception of a critical behavior in the seismological phenomenon and that there are no differentiation between the spatiotemporal statistical features of earthquakes, whether small or large in size or magnitude, and independently of the tectonic environments. It means that, regardless of the physical mechanism of energy release and the consequent emergence of an earthquake, that mechanism operates similarly on all scales of space and time. Moreover, the presence of *q-exponential* distributions and the ability of a *small-world-like* OFC model to reproduce spatiotemporal features of real worldwide earthquakes indicate self-organized criticality and long-range spatiotemporal correlations between seismic events. Recently, we have published the results obtained in the present work in the Chaos, Solitons & Fractals journal [14].

References

[1] Z. Olami, H. J. S. Feder, K. Christensen, *Phys. Rev. Lett.* 68, 1244 (1992).

[2] D. S. R. Ferreira *et al., Phys. Lett. A* 379, 669 (2015).

[3] F. Caruso *et al*., *Eur. Phys. J. B* 50, 243 (2006).

[4] D. J. Watts, S. H. Strogatz, *Nature* 6684, 440 (1998).

[5] C. Tsallis, *J. Stat. Phys.* 52, 479 (1988).

[6] S. Abe, N. Suzuki, *J. Geophys. Res.* 108, 2113 (2003).

[7] A. H. Darooneh, C. Dadashinia*, Physica A* 387, 3647 (2008).

[8] G. Papadakis, F. Vallianatos, P. Sammonds*, Tectonophysics* 608, 1037 (2013).

[9] P. Bak *et al., Phys. Rev. Lett.* 88, 178501 (2002).

[10] J. Davidsen, M. Paczuski, *Phys. Rev. Lett.* 94, 048501 (2005).

[11] A. Corral, *Phys. Rev. E* 68, 035102 (2003).

[12] C. Frohlich, *Annu. Rev. Earth Planet. Sci.* 17, 227 (1989).

[13] C. Frohlich, *Cambridge University Press* (2006).

[14] D. S. R. Ferreira *et al.*, *Chaos, Solitons & Fractals*, 165, 112814 (2022).

Spreading processes in intermittent networks

Juliane T. de Moraes and Silvio C. Ferreira
Universidade Federal de Viçosa, Brazil

Keywords: Networks, Spreading phenomena, Epidemics, Renewal processes.

Many efforts have been devoted to understand spreading phenomena [1]. The importance of developing methods and models in epidemics becomes more evident with the SARS-CoV-2 pandemic. These models can improve health strategies and help the forecast of new outbreaks. In addition, studies of dissemination processes can also be applied to social dynamics [2].

Spreading processes are commonly related to social interactions and can be investigated considering complex networks [3, 4] as underlying substrates for dissemination, in which nodes represents the agents and links the interactions between them. An approach can be derived by picturing the following situation - considering a period of time, one individual has a given number of acquaintances. However, the contacts are intermittent. Therefore, we investigate spreading processes on intermittent networks, in which nodes (or links) become active or inactive across the time. In Fig. 1, an illustration of the intermittent dynamics is shown for nodes and links.

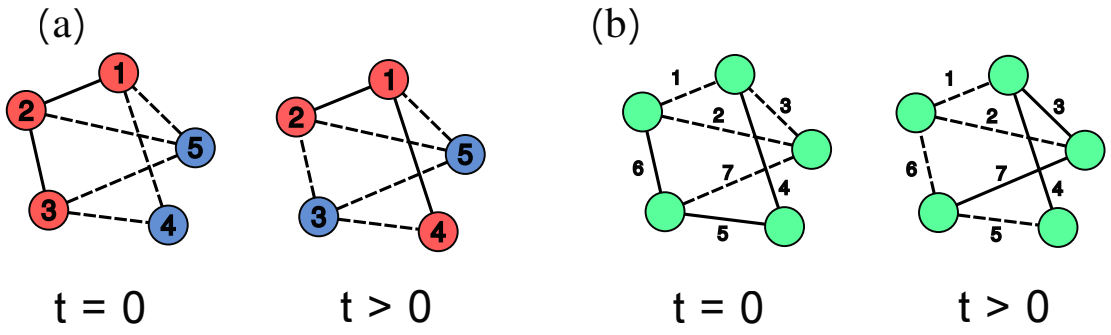


Figure 1: Illustration of the intermittency applied to static networks in different times. (a) Node intermittency, where nodes in red are active and in blue are inactive. (b) Link intermittency, in which links can be active (solid line) or inactive (dashed line).

In our simulations we start building a network structure for an initial condition in which some nodes (or links) are active and the remaining are inactive. The states of each node alternate between active and inactive following independent random processes with inter event times obeying a given probability distribution. If this distribution is exponential, these events are Poisson processes [4]. However, in a more general case, the inter-event time distribution can be chosen, characterizing a general renewal process. For our analysis, we used two classes of inter-event time distributions: exponential $\psi(\tau) \sim e^{-\alpha\tau}$ and power law $\psi(\tau) \sim \tau^{-\nu}$.

We used the uncorrelated configuration model to generate underlying power law degree distributed networks without degree correlations [5]. We observe that the network remains uncorrelated over time. This can be understood by the fact that the intermittency maintains the number of active nodes (or links) approximately constant depending on the activity and inactivity probability distributions. This behavior occurs for the two types of inter-event time distributions investigated.

In the SIS model an individual (or node) can be Susceptible and become Infected upon contact with another infected node with a given infection rate λ. Also, it can heal and return to the Susceptible state at other given rate [6]. This model, although quite simple, is known to undergo a phase transition between an inactive and an endemic (active) regime as the infection rate λ is varied. A method to estimate the critical value of λ is the susceptibility χ, which presents a peak near the epidemic threshold λ_c [7]. In Fig. 3, the susceptibility as a function of infection rate for a static network is compared with the ones obtained with intermittency. As

expected, the node intermittency shows a stronger effect, increasing the epidemic threshold, since the contacts are more limited than in the link intermittency. Our forthcoming analysis includes to increase the size and time of simulation and to apply the intermittency for other network substrates. Also, we intend to study the phase transitions and scaling properties of these networks.

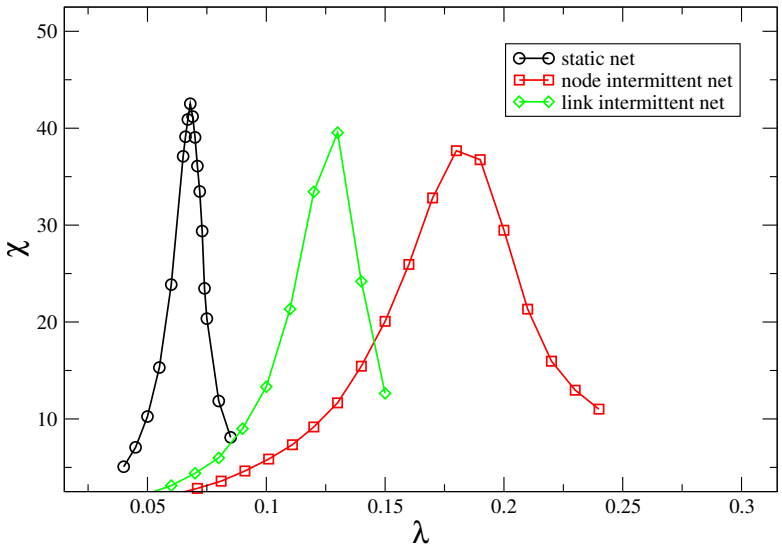


Figure 2: Susceptibility as a function of infection rate for power law degree distributed networks with exponent $\gamma = 2.75$ and size $N = 10^4$. Black symbols are related to the static network, without intermittency. Green and red symbols refer, respectively, to link and node intermittency for the underlying network using a exponential inter-event time distribution with parameter $\alpha = 0.5$

Financial support: CAPES, CNPq, FAPEMIG and INCTSC .

References

[1] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in humans and animals*. Princeton University Press, 2008.

[2] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," *EPJ Data Science*, vol. 8, no. 1, 2019.

[3] A.-L. Barabási, *Network science*. Cambridge University Press, 2016.

[4] N. Masuda and R. Lambiotte, *A guide to temporal networks*. No. vol. 4 in Series on complexity science, New Jersey: World Scientific, 2016.

[5] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, "Generation of uncorrelated random scale-free networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 71, no. 2, pp. 1–4, 2005.

[6] W. Cota, A. S. Mata, and S. C. Ferreira, "Robustness and fragility of the susceptible-infected-susceptible epidemic models on complex networks," *Physical Review E*, vol. 98, no. 1, pp. 1–12, 2018.

[7] S. C. Ferreira, C. Castellano, and R. Pastor-Satorras, "Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 86, no. 4, pp. 1–8, 2012.

Study of patient transfers during the COVID-19 pandemic using complex networks

T. Cicchini [*1], L. Otero[2], A. Salgado[1], A. Yacobitti[2], V. Doldan[2], S. Kochen[2,3], L. Boechi[1], and I. Caridi[1]

[1]Instituto del Cálculo (IC), UBA-CONICET. Buenos Aires, Argentina.
[2]Hospital de Alta Complejidad Nestor Kirchner, El Cruce, Pcia Buenos Aires, Argentina
[3]Unidad Ejecutora de Estudios en Neurociencia y Sistemas Complejos (ENyS, CONICET-HEC-UNAJ)

Abstract

The COVID-19 pandemic placed health systems worldwide in crisis, and the pre-existing health infrastructure had to adapt to its new and changeable conditions. Here we analyze a hospitals network representing the *Red Sudeste* in Buenos Aires, Argentina. This network has health centers from the Red Sudeste as nodes, and each link represents the number of patient transfers between pairs of health centers. A fragmented structure arises when exploring the aggregated network in time. The evolution of the network, separated into three different stages, shows an increase in the efficiency of patient transfers through time.

Keywords: Network Medicine, Temporal Networks, Health Systems, COVID-19

Introduction

The COVID-19 pandemic placed health systems worldwide in crisis. The lack of health personnel and supplies, and the saturation of hospital beds, were some of the effects generated at the beginning of the pandemic. The functioning of hospitals in a network is a common practice that allows the optimization of physical and human resources at critical moments, facilitating the referral of patients of medium and high complexity between the network's hospitals [1]. However, the pandemic generated abrupt changes in the coordination of these networks. Here we consider the *Red Sudeste*, a hospitals network including the Public Hospitals of four municipalities in the Southeast of the Metropolitan Area of Buenos Aires, Argentina. The public hospitals of *Red Sudeste* have different capacities and complexities. In particular, *UPAs*(Promt Care Units), by its initials in Spanish, are primary attention centers and *Módulos Hospitalarios*(Hospital Module) are high complexity health centers. We analyze an aspect of the coordination between the health centers: the patient transference within the network's fourteen hospitals, in the period comprised between 01/03/20 and 15/07/21.

A computer system built at the beginning of the pandemic to collect, organize and display information on the use and availability of patient beds in the hospital network [2, 3] stored all data on transfers within *Red Sudeste*. Representing patient transfers as directed and temporary links between health centers, we study the network's structure change over time. At the same time, we quantify the complexity of the transfers in terms of bed type, clinical risk, and differences between the involved health centers, seeking to understand the functioning of the network at the different key moments of the pandemic.

*tcicchini@df.uba.ar

Hospitals Network: Construction and Aggregated Analysis

From the data collected by the system mentioned above, it is possible to build an origin-destination matrix that accounts for the patient transfers between different health centers. Considering all the data of the studied period, we assemble a weighted and directed complex network, where the nodes represent the hospitals and the weight of the links accounts for the number of transfers between two hospitals. This representation (see, for example, [4]) allows us to study characteristics of the hospital network related to its level of centralization, fragmentation, and efficiency.

After removing links representing less than 4 transfers, the remaining links only connect hospitals from the same district (Figure 1[A]). *El Cruce* is the only exception, connecting to hospitals from different districts. In turn, *Evita Pueblo* and *UPA 10-BE* are isolated from the rest of the network.

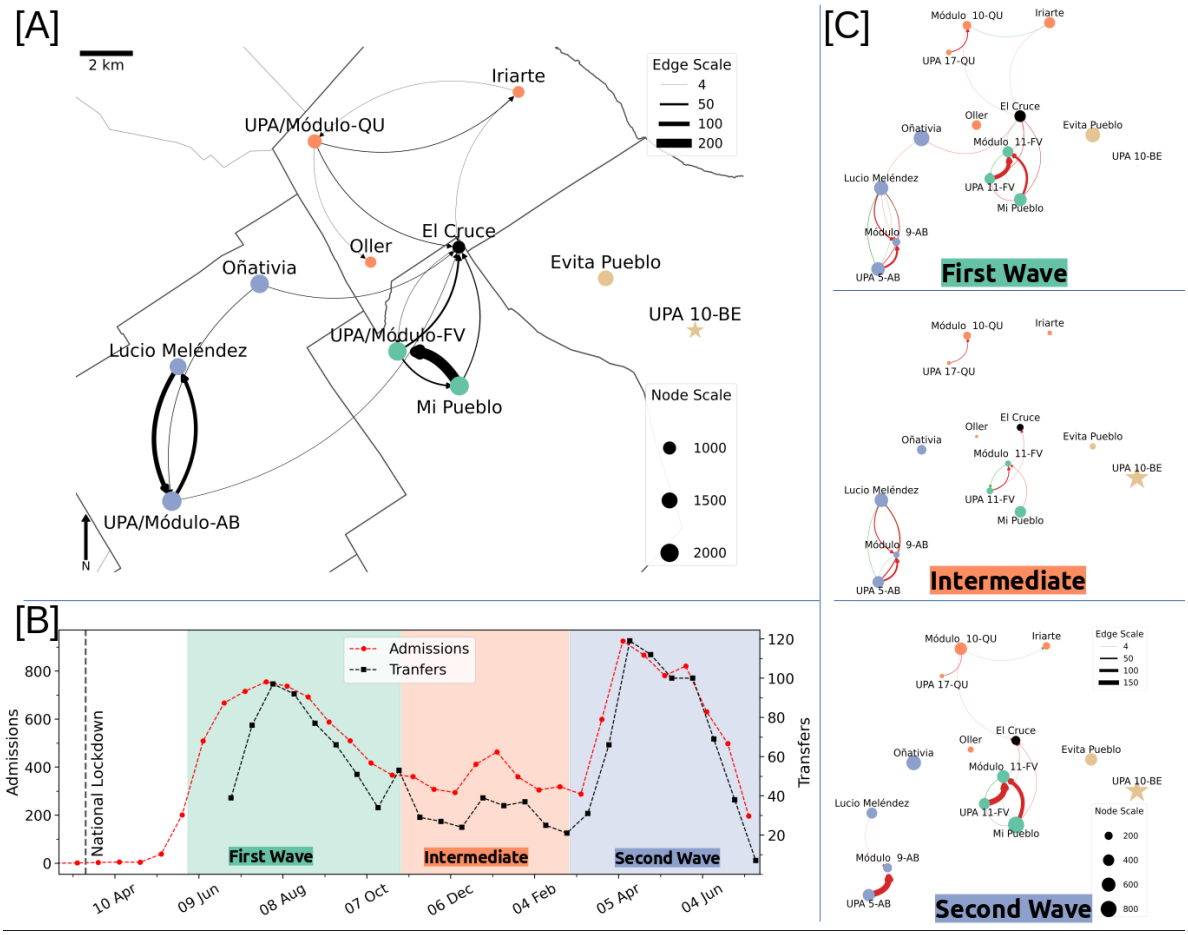


Figure 1: *[A] Aggregated hospitals network. Node size represents the total amount of admissions over the total period. Node layouts correspond to the geographic place of the hospitals. Node colors refer to the distinct districts where the hospitals belong. Edge width linearly stands for the number of transfers between hospitals. [B] Temporal evolution of the total number of transfers and admissions in the different pandemic stages. [C] Temporal snapshots of the temporal hospitals' network.*

Hospital Network: Evolution Throughout the Pandemic

The daily number of patient admissions and patient transfers does not remain constant throughout the pandemic. Moreover, the dynamic of the pandemic changed over time. We divide the period into three different moments, following the criterion defined by the health system administration: first wave (01/06/20 to 31/10/20), intermediate (1/11/20 to 28/02/21) and second wave (1/03/21 to 30/06/2021) (Figure 1[B]). Taking into account the date of each patient transfer, it is possible to build a separate network for each pandemic stage.

	First Wave	Intermediate	Second Wave
Admissions	5017	2674	5559
Total transfers	531	263	612
Ambulance transfers	279 (52.5%)	138 (52.5%)	188(30.7%)
Total traveled distance / median distance by transfer [km]	1518 / 5.41	840 / 5.59	811 / 3.18
UPA-Modulo transfers	252 (47.5%)	123 (46.8%)	424 (69.3%)

Table 1: *Patient admissions and transfers data of the different studied stages. Percentage of Ambulance and UPA-Modulo transfers indicate percentage of total transfers.*

Figure 1[C] shows the networks for each stage. Unlike the aggregated network, the nodes corresponding to *Modulos Hospitalarios* and UPAs were split. While they belong to the same physical space, they operate as different health centers in practice. This distinction allows us to observe the large number of transfers between them. These transfers were made without the need to use ambulances. When observing Figure 1[B] and Table 1, it is evident that the number of admissions and transfers are similar in the first and second waves. However, the UPA-Modulo transfers increase from first to the second wave (see Table 1). Moreover, while the percentage of ambulance transfers represents around 50% of the total transfers on the first wave, they represent only 30% on the second wave (Table 1). In addition, we see an abrupt drop in the total travelled distance and median traveled distance from first to second wave. This reflects the increase in efficiency of the network throughout the pandemic.

Discussion and Further Analysis

The methodology allows us to build a network between hospitals to analyze aspects of the functioning of the Red Sudeste through their transfers. Moreover, by using the temporal characterization of the transfers, we can study the evolution of the network through time. These preliminary results contribute to our better understanding of the behavior of the *Red Sudeste* throughout the COVID-19 pandemic. In particular, the aggregated network analysis shows that the system is highly fragmented with *El Cruce* being the node that connects health centers from different districts. This makes sense because *El Cruce* Hospital has the greatest level of complexity in *Red Sudeste*, receiving patients with severe risk from many health centers. Also, the analysis by stages confirms that transfers became more efficient, since ambulance transfers and traveled distance dropped significantly.

Currently, our work focuses on the different levels of complexity that characterize hospitals (according to their care capacities and resources) and transfers (considering the type of beds and the patient's state of health). Making use of this information will provide a deeper understanding of the system.

References

[1] Redes y Territorios: aportes para planificar la política de salud en nuestra región, Daniela Alvarez, Magali Turkenich, Universidad Nacional Arturo Jaureche (2020)

[2] Yacobitti, A et al. "Clinical characteristics of vulnerable populations hospitalized and diagnosed with COVID-19 in Buenos Aires, Argentina." Scientific reports vol. 11,1 9679. 6 May. 2021, doi:10.1038/s41598-021-87552-w

[3] https://www.ic.fcen.uba.ar/institucional/herramientas/hospitales-en-red

[4] Kohler, K., Jankowski, M.D., Bashford, T. et al. Using network analysis to model the effects of the SARS Cov2 pandemic on acute patient care within a healthcare system. Scientific Reports 12, 10050 (2022).

Tailoring Benchmark Graphs to Real-World Networks for Improved Prediction of Community Detection Performance

Catherine Schwartz[1,2], Cetin Savkli[1], Amanda Galante[1], and Wojciech Czaja[2]
[1] *Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA and*
[2] *University of Maryland, Department of Mathematics, College Park, Maryland 20742, USA*

We introduce a new methodology for improving the understanding of how different community detection methods are expected to perform on a specific real-world network of interest. A common approach used to compare the performance of community detection methods is to measure their ability to detect ground truth communities in benchmark graphs. Studies that employ this approach are typically based on general benchmark model parameters that are selected to create benchmark graphs with realistic community structure. The authors of the studies then provide guidance on how to choose an appropriate method based on the performance results on the generated benchmark graphs. Unfortunately, researchers and practitioners may follow the guidance even though the real-world network they are working with may not have any resemblance to the benchmark graphs used in the study. For example, their real-world network may be highly connected, whereas the general benchmark model parameters may create graphs that are not. In this study, we demonstrate that, by running experiments on tailored benchmark graphs where model parameters are chosen to match a specific real-world network of interest as closely as possible, researchers can obtain a better understanding on how well community detection methods will work on that particular network. Since the popular LFR benchmark [1] (See FIG. 1a) cannot create certain community structures seen in real-world networks, we motivate researchers to additionally consider a recently proposed network growth model called the nPSO benchmark [2] (See FIG. 1b) when determining which benchmark model to use for their tailored benchmark graphs. For this study, our first real-world network of interest was the email-Eu-core network, a publicly available dataset from SNAP [3] which we reveal has a community structure similar to the community structure generated by the nPSO benchmark. We show that the performance of the community detection methods on the email-Eu-core network is highly correlated with the performance of the same methods on the corresponding tailored benchmark graphs (r = 0.93, FIG. 2a). This suggests that methods that performed well on the tailored benchmark graphs also performed well on the real-world network. Conversely, the performance of community detection methods on the email-Eu-core network was not correlated with the performance of the same methods on unrelated benchmark graphs (r = -0.25, FIG. 2b), meaning the methods that performed well on the unrelated benchmark graphs did not perform well on the email-Eu-core network. We further demonstrate how to create tailored benchmark graphs when a real-world network has no associated ground truth and introduce a tool that can be used to help ensure the appropriate community structure is reflected in the tailored benchmark graphs. We use another publicly available real-world network from SNAP [3], the DBLP collaboration network, that does not have non-overlapping ground truth communties and we use the mentioned tool to illustrate that the community structure of the DBLP network is similar to the community structure generated by the LFR benchmark. There are a number of similarities between how the community detection methods perform on the tailored benchmark graphs for the DBLP network and how the methods perform on the actual network itself. The results inform the type of errors the community detection methods are expected to make and identify which methods will tend to overpredict or underpredict the number of communities [4]. By utilizing tailored benchmark graphs, researchers and practitioners can select an appropriate community detection method in a more systematic way for a specific network they are studying and reach a better understanding of communities that are generated. This approach will increase trust and confidence in the resulting communities, which is particularly important if the communities are going to be used for downstream analyses.

Keywords – community structure, community detection, benchmark graphs, network models

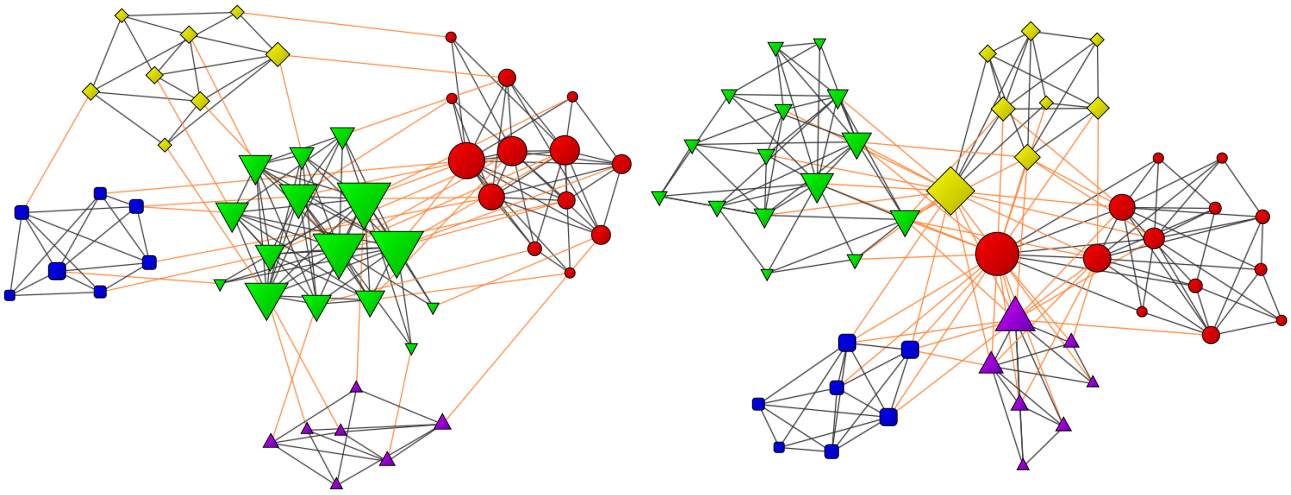
---

[1] A. Lancichinetti, S. Fortunato, and F. Radicchi, Benchmark graphs for testing community detection algorithms, Physical Review E **78** (2008).

[2] A. Muscoloni and C. V. Cannistraci, A nonuniform popularity-similarity optimization (npso) model to efficiently generate realistic complex networks with communities, New Journal of Physics **20** (2018).

[3] J. Leskovec and A. Krevl, SNAP Datasets: Stanford large network dataset collection, `http://snap.stanford.edu/data` (2014).

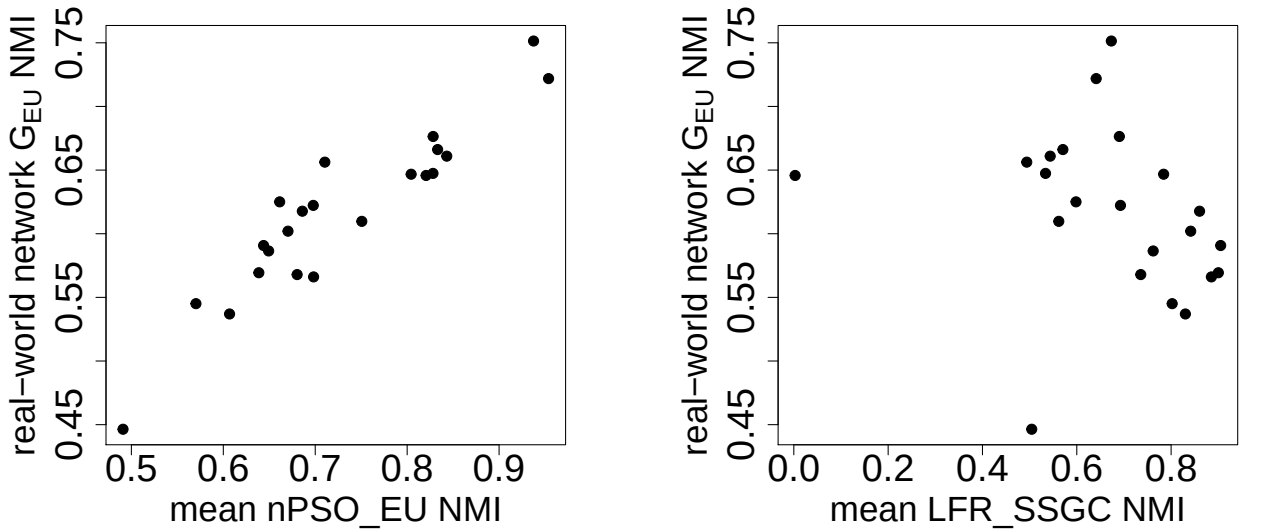
[4] C. Schwartz, *Analyzing Semi-Local Link Cohesion to Detect Communities and Anomalies in Complex Networks*, Ph.D. thesis (2021).



(a) Degree assortative community structure created with the LFR benchmark [1]

(b) Hierarchical community structure created with the nPSO benchmark [2]

FIG. 1: Example graphs with five communities. Color and shape indicates a vertex's community. Vertex size corresponds to its degree. Internal edges are black and external edges are orange.



(a) nPSO_EU are tailored nPSO benchmark graphs [2] for the G_{EU}, the email-Eu-core network

(b) LFR_SSGC are unrelated LFR benchmark graphs [1] for the G_{EU}, the email-Eu-core network

FIG. 2: Comparison of the mean normalized mutual information (NMI) from benchmark graphs to the NMI from G_{EU}, the email-Eu-core network [3], of different the community detection methods. Each point represents a different community detection method.

The Complex Network Analysis of Power Grid: A Case Study of the Argentinian Network and its Vulnerability

Ayelen Bargados[*] and Viktoriya Semeshenko[**]
[*]Fundación Observatorio PyME. Universidad de Buenos Aires.
[**]Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina.

Keywords: Power grid | Complex Networks | Avalanches | Simulations

The communication, transport and energy infrastructure are essential for everyday life. The effects of damage and/or infrastructure deficits, if it happens, are far-reaching. In the case of electric supply power, the occurrence of failures can cause interruptions in the communication channels and loss of information, alterations in transport, problems with the banking circuit and the operation of the capital market, and affect the residential activity as well as manufacturing processes, mentioning these as few examples based on what has been seen in successive blackouts that took place in different parts of the world.

In this work we analyze the impact of cascade failures (which potentially produce "blackouts" (Albert et al., 2004)) in the Argentinian power grid system, from the perspective of a network theory approach that allows us to study the structural characteristics of this network and its behavior under shocks (Albert & Barabási (2002); Boccaletti (2006); Pagani & Aiello (2013)). This research tries to account for characteristics that haven't been studied previously in the Argentinian electric power transmission system. Energy infrastructure and, in particular, the lay-out of transmission networks, can contribute to territorial cohesion and, therefore, to socio-economic development. In particular, it is interesting to study the network of connections of isolated systems that initially tried to respond to the demands of each territory of the country, according to the development of social and economic sectors that took place in each area, and on the particular geography, where the large centers of electricity consumption resulted to be far from the most important centers of generation.

The national power grid system is, in turn, segmented into three component subsectors, well differentiated in terms of the activities that concern each one (and in regulatory matters): generation, transport/transmission, and distribution. Given the regulation conditions, the macroeconomic context and the microeconomic behaviors, the sector has reached a critical state in terms of provision of electrical energy to the final user. During the last decade, the deterioration of service quality and supply interruptions (some of great magnitude) have become increasingly frequent and widespread, following multiple factors such as the growth of urban demand, some climatic phenomena and the absence of sufficient investments. These events imply economic costs that could more than compensate for the potential benefits of existing electrical infrastructure lay-out.

In this work, we are using tools and modeling techniques coming from network theory to characterize the Argentinian power grid network and understand its behavior and robustness. In particular, we want to reveal specific topological and dynamic characteristics of the network that allows us to recognize the origin of failures and help to design policy making in order to prevent future blackouts.

The networks under study correspond to the high-voltage transmission systems, which transport the electrical energy from the generators to the distribution heads (that are also denominated substations). The Argentine electrical network is studied at two

geographical levels: the national system SADI[1] and the regional system GCABA[2]. First, we start with data[3] collection. Initially, the information was available in terms of geographic map formats only. These maps were digitized, from which the nodes and main connections have been extracted in stylized form. The constructed networks replicate the real power grids, see Fig. 1. We then characterized the topologies of these networks and their main descriptive statistical properties.

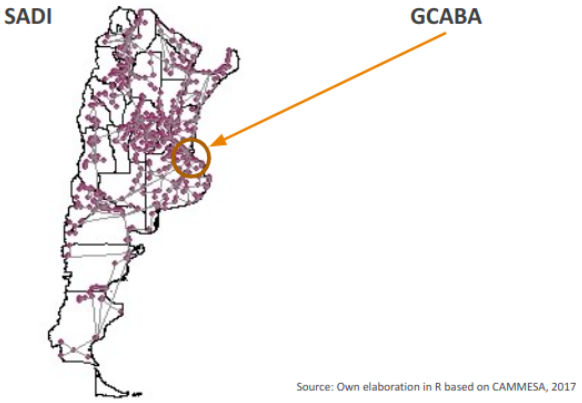


Fig. 1 Network representation of the power grids under study: SADI, GCABA.

Ensuing, we performed simulations to reveal the network robustness under shocks (disturbances due to different causes have been discussed in Crucitti et at. (2004); Kinney et al (2005); Martins et al (2016); Motter (2004); Rosas-Casals et al (2007); Saniee Monfared et al (2014)). The dynamics of the networks against exogenous shocks is simulated under two scenarios: random shocks (connections are broken at random) and directed, or preferential shocks, perturbing certain selected nodes according to different criteria of relevance. Likewise, the global state of the networks is analyzed sequentially as a greater proportion of the connections suffer failures, focusing the study on events of type "blackout" –cascade failures–, (see Fig.2 for a particular example).

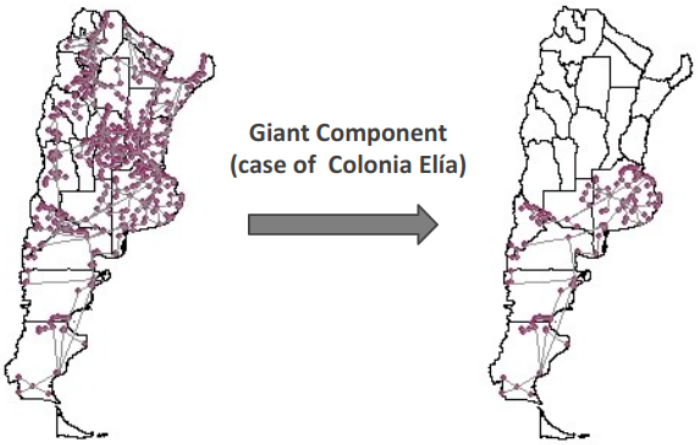


Fig. 2 FATHER'S DAY BLACKOUT
On June 16, 2019 in Argentina (Father's Day), a blackout took place that affected a big part of the country. We simulate the shock, retiring the nodes whose links were involved in the real failure, and reach similar results to the observed effects. We show the geo-representation of the giant component, once one of the nodes involved in the blackout origin is out of service.

[1] SADI stands for the acronym in *Spanish* of Interconnection Argentinian System.
[2] GCABA stands for Autonomous City of Buenos Aires and Greater Buenos Aires.
[3] The data used consists of the geographical layout of generators and transformer stations of the electric power transmission system throughout the country, and for the GCABA power grid there is also the geographical arrangement of the distribution substations. This information is available in the Electricity Wholesale Market Management Company and in the National Ministry of Energy. First, it was necessary to transform the geographic maps into spreadsheet inputs, with tabulation according to the way in which the information is processed by the software used for network analysis in this research. Likewise, this information was enriched through exchanges with experts in the field (engineers and technicians). For the purposes of the construction and graphic presentation of the power grids, information from the IDE Conurbano of the National University of General Sarmiento, National Geographic Institute and GADM maps and data is also incorporated.

In terms of the main results, the networks present properties of small worlds (Watts (2004); Latora & Marchiori (2003)), where energy is transmitted more "quickly" but where shocks also expand more easily. In this sense, the results would indicate that the removal of random elements causes relatively minor damage to the removal of main nodes: when a minority of central nodes are out of service, a tendency to collapse of practically the entire network is observed, unlike the removal of a similar (or even higher) proportion of nodes at random, when the damage can be much less. That is, the networks under study seem to behave in a robust manner in the face of random shock while they are vulnerable to the intentional removal of central nodes (topologically defined).

References

1. Albert, R., Albert, I., & Nakarado, G. L. (2004). Structural vulnerability of the North American power grid. *Physical Review E*, *69*(2), 1–10. https://doi.org/10.1103/PhysRevE.69.025103
2. Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47
3. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4–5), 175–308. https://doi.org/10.1016/j.physrep.2005.10.009
4. Crucitti, P., Latora, V., & Marchiori, M. (2004). A model for cascading failures in complex networks. *Physical Review E*, *69*(4), 045104. https://doi.org/10.1103/PhysRevE.69.045104
5. Kinney, R., Crucitti, P., Albert, R., & Latora, V. (2005). Modeling cascading failures in the North American power grid. *European Physical Journal B*, *46*(1), 101–107. https://doi.org/10.1140/epjb/e2005-00237-9
6. Latora, V., & Marchiori, M. (2003). Economic Small-World Behavior in Weighted Networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, *32*(2), 249–263. https://doi.org/https://doi.org/10.1140/epjb/e2003-00095-5
7. Martins, G. C., Oliveira, L. S., Ribeiro, F. L., & Forgerini, F. L. (2016). Complex Network Analysis of Brazilian Power Grid. *Scientia Plena*, *14*(10), 1–7. Retrieved from http://arxiv.org/abs/1608.07535
8. Motter, A. E. (2004). Cascade control and defense in complex networks. *Physical Review Letters*, *93*(9), 098701. https://doi.org/10.1103/PhysRevLett.93.098701
9. Pagani, G. A., & Aiello, M. (2013). The Power Grid as a complex network: A survey. *Physica A: Statistical Mechanics and Its Applications*, *392*(11), 2688–2700. https://doi.org/10.1016/j.physa.2013.01.023
10. Rosas-Casals, M., Valverde, S., & Sole, R. V. (2007). Topological Vulnerability of the European Power Grid Under Errors and Attacks. *International Journal of Bifurcation and Chaos*, *17*(07), 2465–2475. https://doi.org/10.1142/S0218127407018531
11. Saniee Monfared, M. A., Jalili, M., & Alipour, Z. (2014). Topology and vulnerability of the Iranian power grid. *Physica A: Statistical Mechanics and Its Applications*, *406*, 24–33. https://doi.org/10.1016/j.physa.2014.03.031
12. Watts, D. J. (2004). *Small Worlds: the dynamics of networks between order and randomness* (8th ed.). Princeton, N.J.: Princeton University Press.

The different effects of non-pharmaceutical interventions in epidemic models based on networks versus mixing matrices

Alberto Aleta[1,4], Adequate Mhlanga[2], Maria Litvinova[2], Marco Ajelli[2,3,*], Yamir Moreno[4,*], and Alessandro Vespignani[3,1,*]

[1]ISI Foundation, Turin, Italy
[2]Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA
[3]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, USA
[4]Institute for Biocomputation and Physics of Complex systems, University of Zaragoza, Spain
[*]Corresponding authors: marco.ajelli@gmail.com, yamir.moreno@gmail.com, alexves@gmail.com

Mathematical models are one of the key tools to combat epidemics of emerging diseases. As the recent COVID-19 pandemic has shown, it is essential to properly model human contact patterns to truly understand the epidemic dynamics. The classical approach is based on mixing matrices that encode the number of interactions between certain groups of the population, being age-mixing matrices the most common of them. More recently, the use of networks has sparked a plethora of discoveries in the field of mathematical epidemiology. Despite their differences, it is generally possible to reproduce the same results with both approaches as long as one properly calibrates them. However, once this had been done, if one introduces non-pharmaceutical interventions aimed at reducing the number of contacts in the population one might get very different outcomes.

In this work, we first build multiplex contact networks using highly detailed sociodemographic data representing the interactions of individuals in schools, workplaces, households, and the general community. Then, we derive the age-mixing matrices that would encode exactly the same set of human interactions from these networks. We implement a classical SIR model, as well as a model for influenza and for COVID-19, using each of these approaches, and calibrate them to yield the same results under a baseline scenario. We aim to understand what would each model predict when a certain intervention is put in place.

In particular, we mimic the closure of schools, workplaces, and the reduction of contacts in the community that has been observed during the COVID-19 pandemic, and study how stringent they have to be to control an outbreak. Our results show that when interactions are encoded in networks, the models predict that a milder intervention is enough to stop an outbreak in comparison to the approach with age-mixing matrices (see figure 1). This has very important policy implications given that the large majority of models that have been used during the COVID-19 pandemic are based on approaches akin to the age-mixing matrix one. However, we acknowledge that building such networks is a very complex process that requires a large amount of data, and the models build on networks are not free from caveats. As such, we also explore the strengths and weaknesses of each approach and give some recommendations regarding the strength of the claims that can be done with one approach or the other.

Keywords: *compartmental models; network epidemiology; multilayer networks*

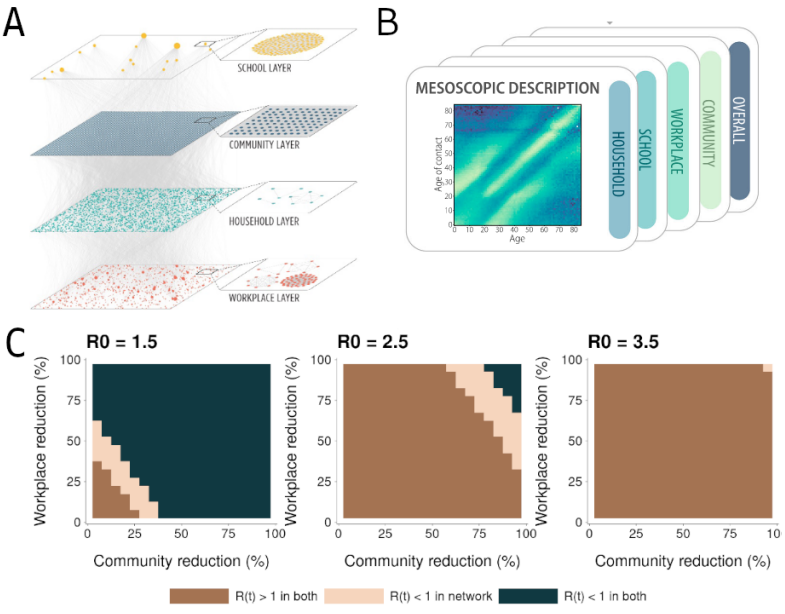


Figure 1: **Modeling non-pharmaceutical interventions**. A: multilayer network representing the daily social interactions in a large population of individuals. B: age-mixing matrices extracted from the aggregation of contacts in each layer. C: Effect of interventions on the spreading of an outbreak. In both models, we implement a SIR model with a recovery period of 2.8 days and explore three values of R_0: 1.5, 2.5 and 3.5. We reduce the number of interactions in the workplace and community layers, and explore the value of $R(t)$ after the intervention. If contacts are modeled using networks, milder interventions can get an outbreak under control in comparison to models based on mixing matrices.

The topology of the skill relatedness network of technologies.

M.Sc. Sergio Palomeque

November 7, 2022

Keywords: skill relatedness; small world; patents technologies

This paper uses the transition of inventors between technologies to estimate their proximity in terms of required skills. This proximity allows the construction of a skill relatedness network between technologies. The aim of the paper is to analyse which topological structures characterise this skill relatedness network.

In recent decades, academic research on the conditions that constrain or boost the generation of new technologies, as an expression of the combination of diverse knowledge and skills, has gained ground (Dosi & Nelson, 2010; Nelson & Winter, 1982). In this literature, a field of research has been developed that interpret the creation of knowledge, in the form of invention of new solutions or combinations, as the product of a complex adaptive system. This system is capable of leading to emergent phenomena (innovations) that transcend the sum of its parts (Fleming & Sorenson, 2001; Frenken, 2006; Martin & Sunley, 2007; Sorenson, Rivkin, & Fleming, 2006).

It is possible to operationalise these ideas by means of the methodological framework that arises from Hidalgo, Klinger, Barabasi, and Hausmann (2007) and Hidalgo and Hausmann (2009), where it is postulated that the economic development of a country is strongly related to the level of complexity of its productive structure. Of particular relevance within this framework of analysis is the "Principle of Relatedness" (Hidalgo et al., 2018), which uses information on the co-occurrence of activities to estimate their proximity.

For the purposes of this research, one particular type of relatedness is relevant, which is called skill relatedness. This approach stems from the work of Neffke and Henning (2013); Neffke, Otto, and Weyh (2017), where the movement of workers between industries is used as an input to estimate skill relatedness. This application of the relatedness principle is appropriate for inferring relatedness of skills between technologies, to the extent that we can observe inventors patenting in one and then in another technology. By this we will say that, if the number of inventors moving from technology X to technology Y is above what we would expect in a probability distribution conditional on the relative sizes of the production of both, then technology Y has a relatedness of skills to technology X. The works that have studied the relatedness between technologies do so by considering co-citation between technologies (Alstott, Triulzi, Yan, & Luo, 2017; Kogler, Rigby, & Tucker, 2013), but there are no works that study the relatedness of skills, between technologies, expressed by the movement of inventors.

This research make use of data from the United States Patent and Trademark Office (USPTO), obtained through the PatentsView project, which provides disambiguated information on the inventors and owners of patents, as well as the technologies in which each patent is classified, since 1976 to 2020. Also the database used allows us to know the order of the technological fields in which the patent is classified, for which the International Patent Classification (IPC) will be used.

With this information it will be possible to construct transition matrices (F) between technologies, based on the movements of inventors at the global level. To do this, if an individual patented in a technology j and then does so in a technology i, in our matrix we will have that $F_{j,i} = 1$. By aggregating all the moves, between each pair of technologies,

we obtain an asymmetric matrix of $N \times N$, where N is the total number of technologies. From the F matrix, we can estimate a null model that allows us to measure the deviation of the observed transitions, with respect to the expected ones ($\hat{F}$). This is the basis for obtaining, following Neffke and Henning (2013) and Neffke et al. (2017), the skill relatedness indicator (sr) which is defined as:

$$\hat{F}_{i,j} = \frac{\sum_j F_{i,j} \sum_i F_{i,j}}{\sum_i \sum_j F_{i,j}}$$

$$sr^{**}_{i,j} = \frac{F_{i,j}}{\hat{F}_{i,j}}$$

This indicator compares the observed flows ($F_{i,j}$) with the estimated ($\hat{F}_{i,j}$). Therefore, if $0 \geq sr^{**}$, the observed flows going from technology i to j are below, or equal to, the estimated ones. Values of $sr** > 1$ indicate that the flow of individuals from i to j exceeds the expected, and can be interpreted as a relatedness of skills between the two technologies. The indicator defined in this way has the disadvantage of having a steep right tail, because it takes values between zero and infinity. Therefore, a normalisation of the indicator as presented in the following equation is proposed, where it is also imposed that each activity is perfectly related to itself.

$$sr^{*}_{i,j} = \begin{cases} \dfrac{sr^{**}_{i,j} - 1}{sr^{**}_{i,j} + 1} & \forall\, i \neq j \\ 1 & \forall\, i = j \end{cases}$$

Finally, the skill relatedness structure is symmetrized as follow:

$$sr_{i,j} = \frac{1}{2}\left(\frac{sr^{*}_{i,j} + sr^{*}_{j,i}}{2} + 1\right) \tag{1}$$

The equation 1 establishes that $sr \in [0,1]$, where pairs of technologies that show values close to 0 will be considered as markedly dissimilar in terms of skill requirements and those that are close to 1 will be those that are closer in this sense. The *SR* matrix, defined by the sr value between each pair of technologies, can be viewed as an undirected, weighted network. This network is the unimodal projection of the bipartite network of co-occurrence of inventors in technologies. In order to dichotomise this network, it is necessary to establish a threshold ($\theta \in (0,1]$) for the variable sr, from which we say that the skills relatedness is significant.

Based on the theoretical approach of this paper, if the SR describes the complex adaptive system that leads to the emergence of new technologies, then we can expect to observe a small-world structure in this network. Therefore the following hypotheses are proposed:

1. Links between knowledge communities allow to traverse the network in a reduced number of steps.
2. Relatedness links are more likely to be established between technologies that have relatedness links in common with other technologies.

The first implies that the diameter of the network is low and decreasing over time. The second hypothesis can be analysed through the evolution of the transitivity of the network, where it is expected to be high and increasing over time.

The formation of links in this network can be interpreted as the establishment of relatedness between two technologies. Understanding this process is a key input for the design of related and unrelated diversification policies, as it provides inputs for policy makers to define diversification strategies that are consistent with available capabilities.

The following figures show preliminary results for some relevant indicators.

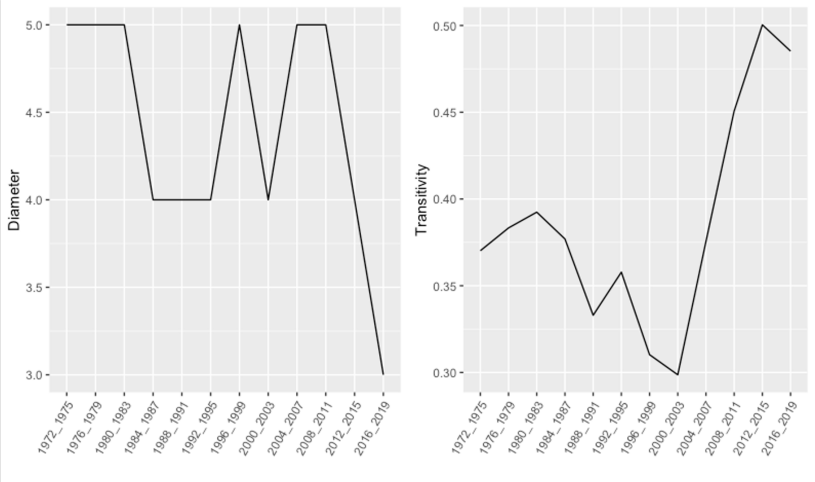


Figure 1: Left: Diameter evolution. Right: Transitivity evolution

References

Alstott, J., Triulzi, G., Yan, B., & Luo, J. (2017). Inventors' explorations across technology domains. *Design Science*, *3*, e20. doi: 10.1017/dsj.2017.21

Dosi, G., & Nelson, R. (2010). Technical Change and Industrial Dynamics as Evolutionary Processes. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (Vol. 1, pp. 51–127). Elsevier B.V. doi: 10.1016/S0169-7218(10)01003-8

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, *30*(7), 1019–1039. doi: 10.1016/S0048-7333(00)00135-9

Frenken, K. (2006). Technological innovation and complexity theory. *Economics of Innovation and New Technology*, *15*(2), 137–155. doi: 10.1080/10438590500141453

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., ... Zhu, S. (2018). The Principle of Relatedness. In *Springer proceedings in complexity* (pp. 451–457). doi: 10.1007/978-3-319-96661-8_46

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575. doi: 10.1073/pnas.0900943106

Hidalgo, C. A., Klinger, B., Barabasi, A.-L. A., & Hausmann, R. (2007). The Product Space Conditions the Development of Nations. *Science*, *317*(5837), 482–487. doi: 10.1126/science.1144581

Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping Knowledge Space and Technological Relatedness in US Cities. *European Planning Studies*, *21*(9), 1374–1391. doi: 10.1080/09654313.2012.755832

Martin, R., & Sunley, P. (2007). Complexity thinking and evolutionary economic geography. *Journal of Economic Geography*, *7*(5), 573–601. doi: 10.1093/jeg/lbm019

Neffke, F., & Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, *34*(3), 297–316. doi: 10.1002/smj.2014

Neffke, F., Otto, A., & Weyh, A. (2017). Inter-industry labor flows. *Journal of Economic Behavior and Organization*, *142*, 275–292. doi: 10.1016/j.jebo.2017.07.003

Nelson, R., & Winter, S. (1982). *An evolutionary theory of economic change*. Harvard Business School Press, Cambridge.

Sorenson, O., Rivkin, J. W., & Fleming, L. (2006). Complexity, networks and knowledge flow. *Research Policy*, *35*(7), 994–1017. doi: 10.1016/j.respol.2006.05.002

Theoretical frameworks for the SIRS model in complex networks with different localization patterns

José Carlos M. Silva,[1] Diogo H. Silva,[2] Francisco A. Rodrigues,[2] and Silvio C. Ferreira[1,3]

[1] *Departamento de Física, Universidade Federal de Viçosa, 36570-900 Viçosa, Minas Gerais, Brazil*

[2] *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP 13566-590, Brazil*

[3] *National Institute of Science and Technology for Complex Systems, 22290-180, Rio de Janeiro, Brazil*

In the present work, we investigate the performance of theoretical approaches in the prediction of properties of the SIRS epidemic model, which involves immunity periods $1/\alpha$ and whose acronym indicates the allowed states (*susceptible*, *infected* and *recovered*) as well as the transitions among them.

As a special case, we have the SIS model ($1/\alpha = 0$), whose activation mechanisms in power-law degree distributed networks, $P(k) \sim k^{-\gamma}$, are well known. They involve self-sustained activity by means of feedback mechanisms in subextensive subgraphs [1], thus spreading to the rest of the network, with the epidemic threshold vanishing in the thermodynamic limit [2, 3]. The *quenched mean-field theory* [4] predicts this behavior qualitatively, with the threshold $\lambda_c = \frac{1}{\Lambda_A}$ vanishing in the thermodynamic limit, where Λ_A is the Largest Eigenvalue (LEV) of the adjacency matrix A_{ij}. An enhancement of the QMF theory is the *pair-quenched mean-field theory* (PQMF), which takes into account the dynamic correlations at a pairwise level [5].

In the SIR limit ($1/\alpha = \infty$), there is a transition to a state in which the fraction of recovered agents is finite, with a vanishing epidemic threshold in power-law distributed networks with $\gamma \leq 3$ and a finite one when $\gamma > 3$. This behavior is well captured by the *heterogeneous mean-field theory* and the exact threshold is predicted by the *message passing* approach [6] on top of tree-like networks.

It is not clear which theoretical approach would be more suitable for the SIRS model with finite immunity periods $1/\alpha$, what is the main concern of our present work [7]. To investigate their properties, we developed PQMF equations for the SIRS model on networks and the rDMP equations reported in [8], whose predictions for threshold and Inverse Participation Ratio (IPR) [9, 10] were compared with extensive stochastic simulations. For more details, we refer the reader to [7].

Our results show that PQMF theory outperforms other approaches in networks without considerable localization effects [7]. Fig. 1 shows how localization affects mean-field theory predictions, by introducing a *hub* in a *random regular* network. PQMF theory undergoes strong localization on the hub and its neighborhood, while rDMP theory shows less localized behavior, agreeing qualitatively with simulation results, specially in dynamics with longer immunity times $1/\alpha$. Other results shown

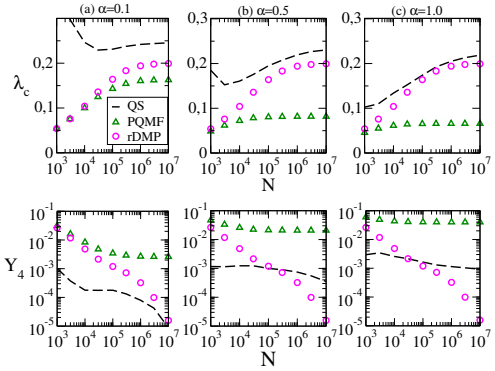


FIG. 1. Finite size scaling for SIRS model in random regular networks ($k_{RR} = 6$) with hub ($k_{hub} = 10^3$). Upper row: threshold λ_c. Lower row: IPR at the threshold. Immunity rates α indicated at the column heading.

in our paper [7] present the same tendency of localization for the PQMF theory in the asymptotic limit, predicting a vanishing epidemic threshold, while simulations and rDMP theory predict delocalization and a finite threshold in power-law degree distributed networks with $\gamma > 3$ [7].

Keywords: Epidemiology - Mathematical models. Networks (Mathematics). SIRS model. Immunity period. Mean-field theory.

ACKNOWLEDGMENTS

The authors thank the financial support given by: *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) and *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG) - Brazil.

[1] C. Castellano and R. Pastor-Satorras, Competing activation mechanisms in epidemics on networks, Scientific Reports **2**, 371 (2012).

[2] M. Boguñá, C. Castellano, and R. Pastor-Satorras, Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks, Phys. Rev. Lett. **111**, 068701 (2013).

[3] S. C. Ferreira, R. S. Sander, and R. Pastor-Satorras, Collective versus hub activation of epidemic phases on networks, Phys. Rev. E **93**, 032314 (2016).

[4] F. Chung, L. Lu, and V. Vu, Spectra of random graphs with given expected degrees, PNAS (2003).

[5] A. S. Mata and S. C. Ferreira, Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks, Europhysics Letters **103**, 48003 (2013).

[6] B. Karrer and M. E. Newman, Message passing approach for general epidemic models, Physical Review E (2010).

[7] J. C. M. Silva, D. H. Silva, F. A. Rodrigues, and S. C. Ferreira, Comparison of theoretical approaches for epidemic processes with waning immunity in complex networks, Phys. Rev. E **106**, 034317 (2022).

[8] M. Shrestha, S. V. Scarpino, and C. Moore, Message-passing approach for recurrent-state epidemic models on networks, Physical Review E **92**, 022821 (2015).

[9] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes, Localization and spreading of diseases in complex networks, Phys. Rev. Lett. **109**, 128702 (2012).

[10] D. H. Silva and S. C. Ferreira, Dissecting localization phenomena of dynamical processes on networks, Journal of Physics: Complexity **2**, 025011 (2021).

Towards Network-Based Planetary Biosignatures: Atmospheric Chemistry as Unipartite, Unweighted, Undirected Networks

Michael L. Wong[1], Anirudh Prabhu[1], Jason Williams[1], Shaunna M. Morrison[1], and Robert M. Hazen[1]

[1] Earth & Planets Laboratory, Carnegie Institution for Science, Washington, DC, 20015, USA.

Keywords: Astrobiology; Chemical Networks; Atmospheric Chemistry; Structural Network Properties

Introduction: Solé and Munteanu (2004) (S&M) first suggested that the chemical reaction network of Earth's atmosphere is topologically distinct from that of other planetary atmospheres. These authors speculated that the uniqueness of Earth's atmospheric network is due to a nonlinear coupling between the biosphere and the atmosphere through the exchange of gases, implying that the network topology of Earth's atmosphere reflects the presence of life. Furthermore, the coevolution of biosphere and atmosphere suggests that the network topologies of atmospheric chemical networks have the potential to serve as an *agnostic* biosignature, because such a description relies less on specific molecules but rather on the nature of the relationships among molecules. However, before atmospheric networks can be used as an astrobiological tool, their analysis requires further development. Here, we build upon S&M's work and explore a more diverse set of atmospheric networks using new graphical representations and topological metrics to classify the network topologies of planetary atmospheres.

Methods: We map the chemical reaction networks of Solar System atmospheres using reaction lists from the Caltech/JPL photochemical model KINETICS (Allen et al., 1981), a versatile and extensively validated code for simulating planetary atmospheric chemistry. Specifically, we analyze chemical reaction networks for: Venus (Zhang et al., 2012), Modern Earth (Yung et al., 2019, 1980), Mars (Nair et al., 1994), early Mars/Earth (Adams et al., 2021), Jupiter (Moses et al., 2005), Titan (Willacy et al., 2016), and Pluto (Wong et al., 2017). Our network visualizations and analyses are performed primarily using NetworkX (Hagberg et al., 2008).

Results: We visualize planetary atmospheres as force-directed unipartite networks, where nodes are chemical species linked by shared reactions. In Figure 1, node color denotes *degree* (the number of links it has) and node size denotes *betweenness centrality* (the number of shortest paths between other nodes in the network that pass through a node). This visualization allows us to qualitatively gauge network characteristics, such as symmetry, "hub" vs. "spoke" nodes, deceptively important nodes (low degree but high centrality), and node distance.

We quantify network structure using a panoply of well-established network metrics including but not limited to: transitivity, degree distribution, centrality distributions, community detection algorithms, and hierarchical clustering. We also compare the metrics of atmospheric networks to equivalent random networks generated using the Erdős-Rényi model (Erdős and Rényi, 1959).

While our modern Earth network does *not* follow a power-law degree distribution (contrasting with S&M's findings), Earth can be distinguished via different metrics. For example, Figure 2 shows that Earth's *degree assortativity*, a measure of whether nodes of similar degree are connected to one another, stands out against various planetary networks and is more similar to certain biological networks. This finding is not simply due to the unique number of nodes and edges in Earth's network; when compared

to their equivalent random networks, Earth's atmospheric network still stands out amongst the other planetary networks in this study.

Discussion: We speculate that, in principle, it may be possible to use the topology of atmospheric chemical reaction networks as a sign of life. It is known that life on Earth exhibits common network structures across all scales, from biochemical to planetary (Kim et al., 2019). The modular hierarchical structure of biochemical networks hints at functionality, robustness, and error tolerance—attributes that would have been selected for via natural selection (Jeong et al., 2001, 2000; Ravasz et al., 2002). Just as the structures of biochemical networks have been honed by evolutionary processes to promote the survival of individual cells, it may be that any prolific, long-lived biosphere will evolve to exhibit persistence-enhancing features in its global-scale chemical networks.

While informative, unipartite graphs offer a *minimal* description of chemical reaction networks because they lack any information about chemical abundances and reaction rates. Hence, we advocate for the use of *weighted* and *directed* network representations, which we plan to pursue in future work. Such representations will incorporate far more information about chemical networks, which are not merely characterized by whether species are present or absent, but also by their abundances and fluxes. If network metrics can robustly group stages of biological evolution across worlds with different geochemical contexts, this may help uncover a possible universal connection between life and planetary complexity and shed light on a theory of life at the planetary scale.

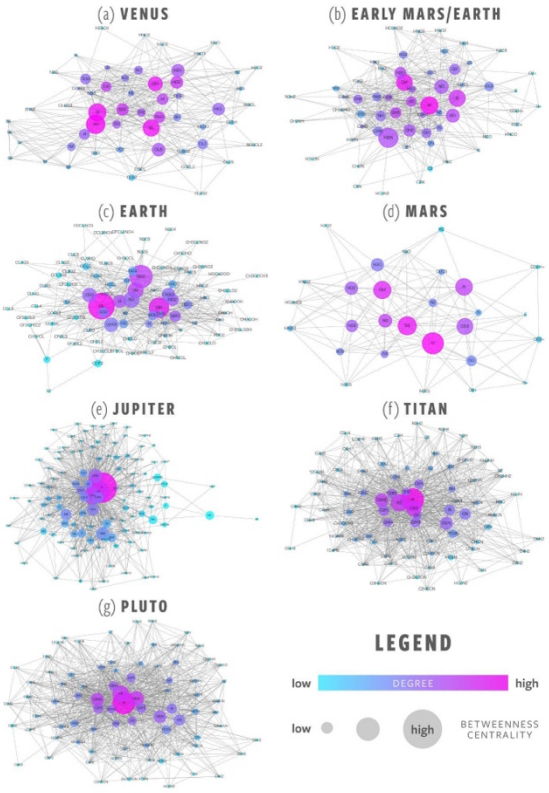


Fig. 1 Solar System atmospheric networks. These diagrams represent the chemistry of planetary atmospheres as force-directed, unipartite, unweighted, undirected graphs. Nodes are colored by their degree and sized by their betweenness centrality.

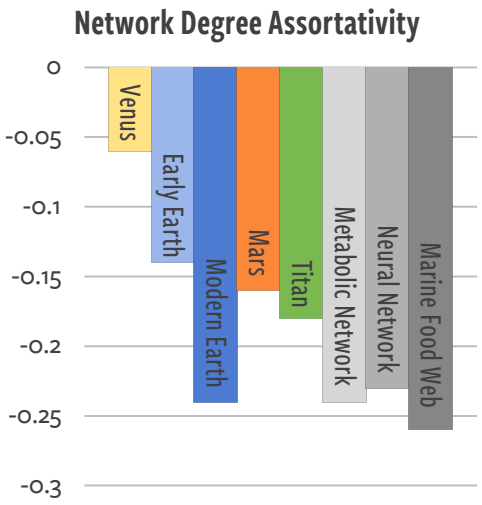


Fig. 2 Network degree assortativity for various planetary atmospheres and biological networks. The lower the assortativity, the more heterogeneous the network is. This is one example of how the topology of modern Earth's chemical network bears resemblance to that of biological networks.

References:

Adams D, Luo Y, Wong ML, et al. Nitrogen Fixation at Early Mars. Astrobiology 2021;21(8):968–980; doi: 10.1089/ast.2020.2273.

Allen M, Yung YL and Waters JW. Vertical Transport and Photochemistry in the Terrestrial Mesosphere and Lower Thermosphere (50-120 Km). J Geophys Res 1981;86(A5):3617–3627; doi: 10.1029/JA086iA05p03617.

Erdős P and Rényi A. On Random Graphs. Publ Math 1959;6(290).

Hagberg AA, Schult DA and Swart PJ. Exploring Network Structure, Dynamics, and Function Using NetworkX. In: Proceedings of the 7th Python in Science Conference. (Varoquaux G, Vaught T, and Millman J. eds) Pasadena, CA USA; 2008; pp. 11–15.

Jeong H, Mason SP, Barabási A-L, et al. Lethality and Centrality in Protein Networks. 2001.

Jeong H, Tombor B, Albert R, et al. The Large-Scale Organization of Metabolic Networks. Nature 2000;407:651–654.

Kim H, Smith HB, Mathis C, et al. Universal Scaling across Biochemical Networks on Earth. 2019.

Moses JI, Fouchet T, Be B, et al. Photochemistry and Diffusion in Jupiter's Stratosphere: Constraints from ISO Observations and Comparisons with Other Giant Planets. J Geophys Res 2005;110:1–45; doi: 10.1029/2005JE002411.

Nair H, Allen M, Anbar AD, et al. A Photochemical Model of the Martian Atmosphere. Icarus 1994;111(1):124–150; doi: 10.1006/icar.1994.1137.

Ravasz E, Somera AL, Mongru DA, et al. Hierarchical Organization of Modularity in Metabolic Networks. Science (1979) 2002;297:1551–1555.

Solé R V. and Munteanu A. The Large-Scale Organization of Chemical Reaction Networks in Astrophysics. Europhysics Letters 2004;68(2):170–176; doi: 10.1209/epl/i2004-10241-3.

Willacy K, Allen M and Yung Y. A NEW ASTROBIOLOGICAL MODEL OF THE ATMOSPHERE OF TITAN. Astrophys J 2016;829(2):1–11; doi: 10.3847/0004-637X/829/2/79.

Wong ML, Fan S, Gao P, et al. The Photochemistry of Pluto's Atmosphere as Illuminated by New Horizons. Icarus 2017;287:110–115; doi: 10.1016/j.icarus.2016.09.028.

Yung YL, Long J, Jiang JH, et al. Effect of the Quasi-Biennial Oscillation on Carbon Monoxide in the Stratosphere. Earth and Space Science 2019;6(7):1273–1283; doi: 10.1029/2018EA000534.

Yung YL, Pinto JP, Watson RT, et al. Atmospheric Bromine and Ozone Perturbations in the Lower Stratosphere. J Atmos Sci 1980;February:339–353.

Zhang X, Liang MC, Mills FP, et al. Sulfur Chemistry in the Middle Atmosphere of Venus. Icarus 2012;217(2):714–739; doi: 10.1016/j.icarus.2011.06.016.

Two-prey-one-predator system: coexistence of sheep, *guanaco*, and puma in the Patagonia region

Jhordan Silveira de Borba[1] **and Sebastián Gonçalves**[1]

[1]Instituto de Física - Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

The modeling of ecological systems, like two or more competing species with or without considering the environment, can help understand the necessary conditions for sustainable coexistence of them. One example of practical interest is in the northern Patagonia of Argentina[1,2]. The *puma* (cougar) and *guanaco* are native species of that region that have coexisted for thousand of years in a predator-prey relation before the introduction of sheep in the XIX century. That changed the Patagonia steppes equilibrium, with the puma shifting to predate preferentially the sheep, while the *guanaco* had to compete with the sheep for grass.

We propose the following set of equations to model a system of three species, being two preys in a hierarchical competition and one predator:

$$\begin{aligned}\dot{x} &= c_x x - \mu_x xz - \frac{x^2}{k_x} - \mu_{xy} xy \\ \dot{y} &= c_y y - \mu_y yz - \frac{y^2}{k_y} \\ \dot{z} &= c_z z(x+y) - e_z z\end{aligned} \tag{1}$$

We have x as the inferior competitor (sheep), y as superior (*guanaco*), and z as the predator (puma). About the constants, c_j controls the population increase of the species j, μ_j tells us about predation of prey j, μ_{xy} is about competition between the preys. k_j is the carrier capacity of the species j, and finally, e_z is the predator extinction rate.

Unlike the *guanaco*, which evolved along with its predator, the sheep is less adapted to flee from the puma, so it is an easier prey. The sheep and *guanaco* also share an almost identical diet, out of 80 plant species, they share 76, which puts the species in direct competition for forage and water. Under natural conditions, the *guanaco* is the superior competitor, better adapted to the local ecosystem; a herd of *guanacos* has the ability to displace a flock of sheep.

The system of differential equations has nine parameters. However, we can take advantage of the fact that dependent variables have no specific meaning. These equations model the changes in the population of each species, but we do not have an exact interpretation of what $x = 1$ means, for example. It can be 1 animal, 1 herd, or up to 1 kg of biomass. So we can do some manipulations to get the following reduced system:

$$\begin{aligned}\dot{x} &= c_x x - \mu_{xy} xy - \mu_x xz - \frac{x^2}{k_x} \\ \dot{y} &= y - yz - \frac{y^2}{k_y} \\ \dot{z} &= z(x+y) - e_z z\end{aligned} \tag{2}$$

The parameters in the equations 2 are not the same as in the equations 1, they are the results of basic operations between the previous variables. We chose to use the same parameters name for notation ease. While there is no absolute meaning for the x, y, z quantities, if any of them is zero, it means that the corresponding species is extinct. Therefore, the most interesting result that the model can show is which species is alive when the system reaches equilibrium, so we will focus on them. A good start is to understand how each variable affects the survival of each species. To investigate it we can use artificial neural networks. The *perceptron* is a basic model of neurons for linear binary classification with supervised learning. The neuron works by assigning a weight to each input, this weight is a measure of the importance of the input to the desired output[3]. We build a simple model where the inputs of each neuron are the 6 independent variables of the system, and the output is the situation of one of the species in the equilibrium state of the system.

After performing a training scheme with 1250 inputs produced via a numerical solution of the system of equations, we performed a validation with 5000 inputs produced in the same way. Each neuron was responsible for predicting whether one

species was alive or not in the steady state, they correctly predicted more than 90% of all situations. The weight of each neuron is illustrated in figure 1.

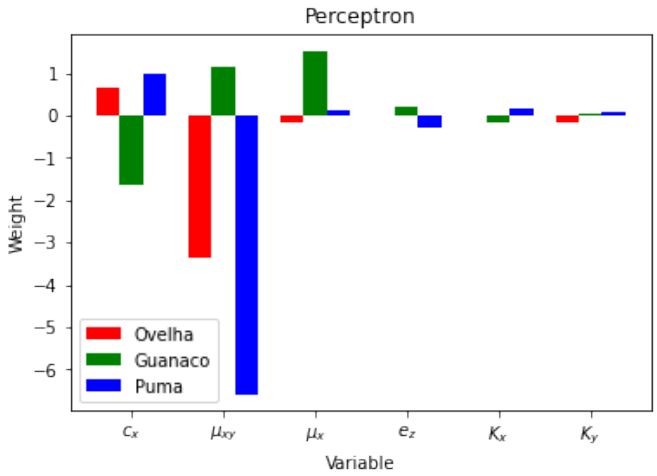


Figure 1. Neuron weights.

Each neuron also has its own variable called bias. Bias tells us the resistance against its activation, and more important than the absolute values of the weights of each input is its ratio against bias. Then, we changed the magnitude of each weight vector to share the same bias for a better comparison.

These results show us that for the range of values we chose to create the dataset, the system is more sensitive to some variables than others, we can also see how each variable affects each species. This helps to get an idea of how we can get from one specific state to another.

A ternary is a graph that allows us to visualize how the system changes according to three variables. If we have 5 variables, we can build 10 different subsets made with 3 different variables each. We can build a graph with 10 ternary graphs arranged in a circle. It will not show us how the system changes according to 5 variables at the same time, but it is difficult to even for us to understand, but it helps us to see how the system changes according to 5 variables in different ways by looking at each ternary graph. We still have one more variable in our system, we can animate the graph using the time to see how the system changes for the last variable.

Returning to the perceptron weights, we can see that if we start from the coexistence of three species, we can reach any other state keeping k_x constant. So we chose to leave k_x to change over time. In figure 2 we can see how the variables are organized in the graph and the final state of the system for different ranges of values.

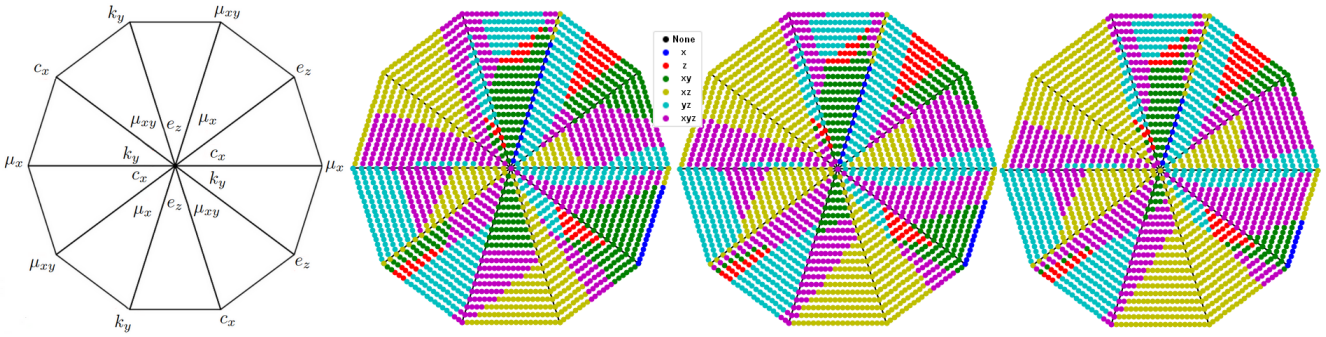


Figure 2. k_x increases from left to right.

We can see that we can reach almost every state by changing μ_{xy}, e_z, and k_y. The only state that is not represented in the graphs is when all are extinct, to reach this result we need to choose a lower value for e_z. This is a counter-intuitive special situation when the predator grows so large that it drives the prey to extinction and then collapses the entire ecosystem. We can see that these 3 variables change the system in different ways. And it is not just about magnitudes, but also, and mainly, whether they contribute to each species' survival or extinction. As a counter-example, we can see that c_x and μ_{xy} affect sheep and puma

in the same way and *guanaco* in the opposite situation, they only have differences in the signal, beyond the magnitudes. Similar situations are for e_z and k_x, μ_x and k_y.

Now we can have a broad understanding of how the system responds to each variable and which path we can choose to take the system from one initial state to another. This is interesting to use as a resource for decision-making.

But if the person behind the decision does not have a certain mathematical background, he can limit himself to the proposed model. Making changes to the equation-based model, solving, and interpreting may require some specific mathematical knowledge.

Agent-based modeling is a different type of modeling where we leverage coding and computing tools. The main idea is that the phenomenon can be modeled using just agents and then writing some simple rules for the interaction between these agents[4]. In this way, we do not model the phenomenon we want to observe directly, it is achieved as an emergent phenomenon of the whole system. And if we want to make some changes to the model, we can do it in an easy way by changing the rules of interactions between agents.

Therefore, we tried to reproduce the results achieved by the system of equations using agent-based modeling. We can define some variables, specifically $k_x = 0.5/c_x$ and $k_y = 0.5$, so we have a probabilistic interpretation of our set of equations that can help us build our agent-based model. The comparison for a specific set of parameters can be seen in figure 3.

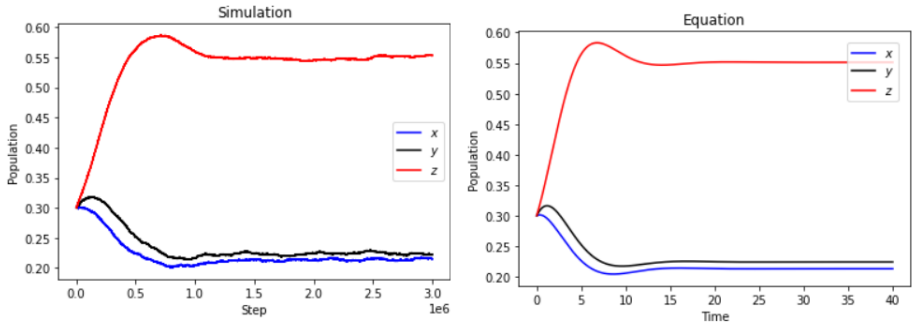


Figure 3. On the left, the simulation of the model, and on the right, the numerical solution of the set of equations.

The similarity is remarkable. In this simple model, each population is represented by an agent with the main attribute we can interpret as the percentage of some total area covered by the animal, and methods are related to reproduction, predation, natural death, and competition. All methods have a one-to-one relationship with the terms of our set of equations. It is easy to think of many ways to improve this model, but this simplicity and strong relationship to the set of equations make it an interesting toy model to explore and a good starting point.

We believe that the agent-based model, with the equation-based model and the tools, exposed earlier to understand and explore the system form an interesting framework to be a "something to think about it", that different researchers can take advantage of and use it as a laboratory to explore some assumptions and ideas in a quantitative way. We hope to contribute to the development of the field and can help in the decision-making process on the management of ecosystems in a sustainable way.

Keywords: mathematical ecology, agent-based model, two-prey-one-predator system, machine learning.

Acknowledgements

Work funded by the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Laguna, M., Abramson, G., Kuperman, M., Lanata, J. & Monjeau, J. Mathematical model of livestock and wildlife: Predation and competition under environmental disturbances. *Ecol. Model.* **309-310**, 110–117, DOI: https://doi.org/10.1016/j.ecolmodel.2015.04.020 (2015).

2. Abramson, G., Laguna, M. F., Kuperman, M. N., Monjeau, A. & Lanata, J. L. On the roles of hunting and habitat size on the extinction of megafauna. *Quat. Int.* **431**, 205–215, DOI: https://doi.org/10.1016/j.quaint.2015.08.043 (2017). Pleistocene human dispersals: Climate, ecology and social behavior.

3. Hagan, M., Demuth, H., Beale, M. & De Jesús, O. *Neural Network Design* (Martin Hagan, 2014).

4. Wilensky, U. & Rand, W. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. The MIT Press (MIT Press, 2015).

Unconsciousness reconfigures modular brain network dynamics

S. M. del Pozo[1,2], H. Laufs[3], V. Bonhomme[4,5,6], S. Laureys[6], P. Balenzuela[1,2], and E. Tagliazucchi[1,2]

[1]Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Buenos Aires, Argentina.
[2]Instituto de Física de Buenos Aires (IFIBA), CONICET. Buenos Aires, Argentina.
[3]Department of Neurology, Christian Albrechts University, Kiel, Germany
[4]Anesthesia and Intensive Care Laboratory, GIGA-Consciousness, GIGA Institute, University of Liège, Liège, Belgium
[5]University Department of Anesthesia and Intensive Care Medicine, Centre Hospitalier Régional de la Citadelle, Liège, Belgium
[6]Department of Anesthesia and Intensive Care Medicine, CHU Liège, Liège, Belgium
[7]Coma Science Group, GIGA Consciousness, University of Liège, Liège,Belgium

Abstract

In this work we used complex network theory to study the dynamics of time-dependent functional brain networks obtained from functional magnetic resonance imaging (fMRI) data during conscious wakefulness and states of reduced consciousness. In order to detect heterogeneous temporal networks communities, we developed a new benchmark to set the optimal parameters of a multilayer modularity maximization algorithm. Then we measured the size and flexibility of the largest multilayer module. We found that unconsciousness reconfigured network flexibility and reduced the size of the largest spatiotemporal module, which we identified with the dynamic core. Our results represent a first characterization of modular brain network dynamics during states of unconsciousness measured with fMRI, adding support to the dynamic core hypothesis of human consciousness.

Keywords: multilayer networks, community detection, fMRI, sleep, consciousness.

Introduction

One of the most influential hypothesis concerning the relationship between conscious experience and neural processes in the human brain is Edelman and Tononi's dynamic core hypothesis [2]. According to this hypothesis, consciousness must be understood as a process that unfolds over time (the "dynamic core") comprising an ever-changing network of regions that exchange information over relatively short time spans. The dynamic core should present a very large number of possible configurations, corresponding to the multitude of available conscious experiences. However, these configurations must also be constrained to represent highly integrated brain states, so the dynamic core consists of a sequence exploring an ample repertoire of highly integrated brain states.

The theory of complex networks provides a framework to directly evaluate the presence of integration and segregation in neuroimaging data acquired during different states of consciousness [3]. A sequence of brain states can be represented as a multilayer network, with each layer encoding transient functional interactions between brain regions during a given time period [4], and the dynamic core can be represented as a time-dependent module evolving in this network. Over the last years, modularity maximization algorithms have been applied to multilayer networks to reveal the rapid and transient structure of whole-brain dynamic networks. However, the relationship between consciousness and the modular structure of multilayer brain networks remains to be investigated.

To clarify this relationship, we constructed multilayer connectivity networks from functional magnetic resonance imaging (fMRI) recordings acquired during the different stages of human non-rapid eye movement (NREM) sleep, and under the effect of propofol, a general anesthetic which increases inhibitory neurotransmission. Our main purpose was to obtain the time-dependent modular structure of these networks using the multilayer Louvain algorithm [9], a method with several free parameters related to the connectivity strength between temporal layers, and the characteristic size of the detected modules. Previous reports using this algorithm either employed an *ad-hoc* choice of parameter values, or performed an exhaustive exploration of parameter space [4, 6]. We introduced a new benchmark for the detection of modules in time-dependent networks with scale-free degree and module size distributions, adapted from a benchmark developed for static networks [5].

After parameter selection using this method, we applied the multilayer Louvain algorithm to obtain the time-dependent modular structure of fMRI functional connectivity networks.

Materials & Methods

Module detection in multilayer networks: We applied a generalized multilayer version of the Louvain algorithm to detect and track modules over time (http://netwiki.amath.unc.edu/GenLouvain/GenLouvain).

Benchmark for time-dependent module detection: We first reproduced a benchmark for static complex modular networks introduced by Lancichinetti et al. [5]. Then we created in this benchmark a temporal evolution through two different dynamic processes adapted from Granell et al., 2015 [8]: merge-split and grow-shrink module dynamics. The combination of these two processes allowed us to represent the most frequent behaviours seen in the dynamics of real modular systems. In Figure 1 **A** the rewiring steps used to generate the dynamics of division and contraction of communities are shown, where the colors of the nodes represent their final communities. Further details can be found in [1].

Node flexibility and the largest multilayer module: We defined two metrics based on the module membership matrix G_{it} given by the multilayer Louvain algorithm. First, we defined the flexibility of a node within a certain multilayer module M, F_i, as the normalized number of times that node entered or left M:

$$F_i = \frac{|\{t : M_{it} \neq M_{it+1}\}|}{T} \tag{1}$$

where M_{it} indicates whether node i at time t belongs to module M, and T is the total number time steps. We computed F_i for the largest multilayer module. We also defined the size of this module (LMM) as the normalized size of the largest module in G_{it},

$$LMM = \frac{\max_i |G_{it}|}{NT} \tag{2}$$

where N is the number of nodes and T the total number of volumes in the recording, so NT is the maximum possible size for the largest module.

More information related to the fMRI data sets, the construction of dynamic networks from fMRI data and statistical analyses are available in [1].

Results

Time-dependent benchmark and parameter selection: We investigated the performance of the multilayer Louvain algorithm [9] based on introducing equally weighted (ω) connections between consecutive temporal layers and equal resolution parameters across layers (γ). Thus, the module detection algorithm depended only on these two parameters. In Figure 1 **B** we introduced a grid of γ and ω values, and for each pair of values we measured the Rand index between heuristic approximations to the ground-truth modules and those detected using the multilayer Louvain algorithm, averaging the results over 500 independent realizations.

The optimal parameters obtained following this procedure were $\gamma = 0.55$ and $\omega = 1$ for both benchmarks (values are indicated as black boxes in Figure 1 **B**). Then, in Figure 1 **C** it is shown the modular structure detected by the multilayer Louvain algorithm using the optimal parameters for merge and split processes. The red lines indicate the expected distribution of module membership labels.

Modular structure of dynamic brain connectivity networks: We applied the multilayer Louvain algorithm using the optimal parameters inferred from the benchmarks to dynamic functional connectivity networks obtained from fMRI data. We computed and compared the flexibility of nodes within the largest multilayer module between wakefulness and each sleep stag. We observed that the majority of nodes decreased their flexibility during sleep, and that regions presenting decreased flexibility during sleep were related to sensory perception, and also included subcortical regions that serve as intermediate stages for the propagation of sensory information towards the cortex, such as the thalamus. We also performed the same analysis and statistical comparison for wakefulness vs. propofol sedation and anesthesia, without finding significant results. Further details can be found in [1].

Finally, we compared the regional probability of belonging to the largest multilayer module in wakefulness vs. sleep, and propofol-induced sedation (S) and loss of consciousness (LOC). Only statistical comparisons between wakefulness, N3 and LOC yielded significant results. Figure 1 **D** presents a comparison of these changes. While changes were more widespread and significant during N3 sleep, LOC was also associated with decreases in sensorimotor regions, and increases in frontal regions.

In Figure 1 **E**, a scatter plot of the change in the probability of belonging to the largest multilayer module for LOC vs. N3 shows that even though less regions were significant for LOC, the pattern of changes was similar to that measured during N3 sleep (R=0.39, p<0.00001). Also, both N3 (0.411±0.011; mean±standard error) and LOC (0.347±0.013) were characterized by smaller sizes of their largest multilayer modules relative to wakefulness (0.446±0.005 and 0.397±0.009 for the sleep and propofol baseline, respectively), as shown in Figure 1 **F**.

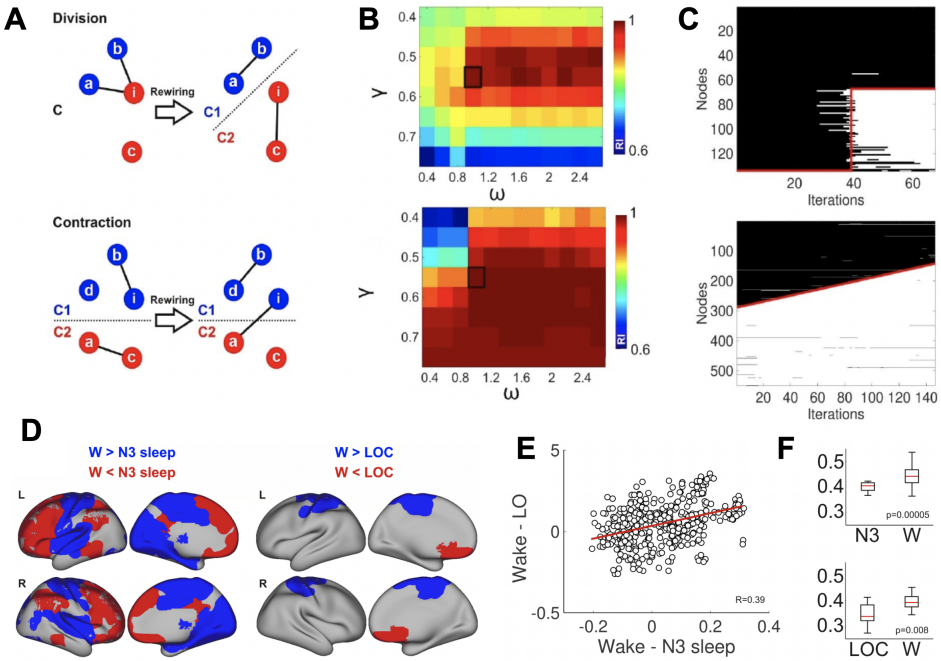


Figure 1: *[**A-C**] Benchmark for time-dependent heterogeneous networks based constructed from two dynamic processes (division and contraction). [**D-F**] Application of the optimized algorithm on data sets.*

Discussion

We investigated for the first time modular brain network dynamics during states of unconsciousness, finding converging evidence of a reconfiguration of the largest multilayer module during deep sleep and general anesthesia. We interpreted these changes in the light of the dynamic core theory, concluding that unconsciousness results in its fragmentation in spite of preserved stability. Future studies should assess whole-brain dynamics simultaneously with different methods to understand whether the dynamic core fluctuates over scales inaccessible to fMRI, and whether these fluctuations are manifest at the behavioral and cognitive levels.

References

[1] del Pozo S. M., Laufs H., Bonhomme V., Laureys S., Balenzuela P. and Tagliazucchi E. Unconsciousness reconfigures modular brain network dynamics *Chaos 31, 093117*, 2021.https://doi.org/10.1063/5.0046047

[2] Edelman, Gerald M and Tononi, Giulio. Reentry and the dynamic core: neural correlates of conscious experience *mit Press Cambridge,Neural correlates of consciousness: Empirical and conceptual questions*, 139 (2000).

[3] Sporns, Olaf. Network attributes for segregation and integration in the human brain *Current opinion in neurobiology Elsevier*, 23, 162 (2013)

[4] Muldoon, Sarah Feldt and Bassett, Danielle S. *Philosophy of Science, University of Chicago Press Chicago, IL* 83, 710 (2016).

[5] Lancichinetti, Andrea and Fortunato, Santo and Radicchi, Filippo. Benchmark graphs for testing community detection algorithms *Physical review E* 78, 046110 (2008)

[6] Bassett, Danielle S and Wymbs, Nicholas F and Porter, Mason A and Mucha, Peter J and Carlson, Jean M and Grafton, Scott T. Dynamic reconfiguration of human brain networks during learning *National Acad Sciences* 19, 566 (2018)

[7] Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing *Wiley Online Library* (1995).

[8] Granell, Clara and Darst, Richard K and Arenas, Alex and Fortunato, Santo and Gómez, Sergio. Benchmark model to assess community structure in evolving networks *Physical Review E* 92, 012805 (2015)

[9] Mucha, Peter J and Richardson, Thomas and Macon, Kevin and Porter, Mason A and Onnela, Jukka-Pekka. Community structure in time-dependent, multiscale, and multiplex networks *American Association for the Advancement of Science,* Science 328, 876 (2010)

Urban segregation patterns for non-homogeneous community linkages

Victoria Arcón[1], Inés Caridi[1], Juan Pablo Pinasco[2] and Pablo Schiaffino[3]

[1] Instituto de Cálculo UBA-CONICET, Argentina
[2] Departamento de Matemática and IMAS UBA-CONICET, Argentina
[3] Departamento de Historia y Estudios Sociales, Universidad Torcuato Di Tella, Argentina

Keywords— ***Computational Social Sciences, Schelling Model, Residential Segregation, Census Data***

INTRODUCTION

Residential segregation is an urban phenomenon in which the population's households are grouped in the physical space according to some distinctive characteristics of the individuals, such as ethnicity, income level, and language, among others. This separation of the different socio-cultural groups in the territory could generate inequalities in access to education, culture, health, and work opportunities. Moreover, the homogenization of the terrain hinders social integration, and it becomes further problematic when associated with stigmatization and discrimination. However, it is also true that being surrounded by similar peers is something positive for creating and sustaining community ties and resolving daily life situations [7].

In this work, we study the mechanisms involved in residential segregation, exploring how the quality and characteristics of the housing location may relate to the importance the individuals give to being surrounded by similar neighbors. We have approached this problem from a social modeling perspective, proposing a variation of the well-known agent-based Schelling's segregation model [8] [9]. This classical model consists of agents of two types arranged in a lattice network that have a preference for being surrounded by some proportion of agents of the same type and they can move from one location of the lattice to another, in order to satisfy their preference. This simple mechanism leads to different segregation patterns, even for mildly discriminatory preferences. An extensive bibliography from economics, mathematics, physics, and computation, contributed to generate many variants of the model [2] [4].

We have introduced non-homogeneous locations in the Schelling model as a weight function over the land (nodes). This function represents objective and subjective valuations of the territory and is related to the relevance that agents give to their community ties. For example, places so prestigious or with full resources and facilities allow individuals to live regardless of their neighbors. Conversely, counting on neighbors becomes crucial under challenging contexts, and ties become relevant to survival. Thus, we define a field on the lattice nodes that modulates the weight of the links.

We study the segregation patterns that arise for different weight functions and show theoretical results that agree with computational simulations [1]. Smooth spatial variations of the weight function, with few minima and maxima, correspond to more significantly segregated neighborhoods with clusters of large scale.

In addition, we study the phenomenon from a data analysis approach, using available Brazilian 2010 census data to visualize and quantify ethnicity-based segregation. Brazil's Sao Paulo and Rio de Janeiro cities allow us to connect the proposed model with the observed patterns. Sao Paulo shows large-scale segregation along all the city and Rio de Janeiro conglomerates near the hills with a higher number of clusters and segregation on a minor scale.

THE WEIGHTED SCHELLING MODEL

We have N agents located as nodes of a connected network Λ_N, without empty sites, so we will denote indistinctly by i, j, ... an agent and its location. A configuration is a function $x : \Lambda_N \to \{-1,1\}$, assigning one of the labels ± 1 to each node (the state of the agent). We have two populations, corresponding to agents of the same state.

A weight function is a bounded function $w : \Lambda_N \to (0,M]$ which assigns a positive real value to each node, bounded above by some $M \in \mathbb{R}$.

We define the neighborhood of a node i as the first neighbors on the network, thus

$$\mathscr{N}_i = \{j \in \Lambda_N \text{ such that } 0 < d(i,j) \leq 1\},$$

where $d(i,j)$ is the usual distance between nodes i and j in a network, is defined as the number of edges of the minimum path between them.

Given a configuration x, we call $U(x,i)$ the *happiness* or utility of an agent located at node i, that depends on the weight w and the configuration x restricted to i and its neighborhood $\mathscr{N}_i$ in the following way:

$$U(x,i) = \frac{1}{(\#\mathscr{N}_i)\, M} \sum_{j \in \mathscr{N}_i} w(j)x(j)x(i),$$

as $\#\mathscr{N}_i$ is the number of nodes in $\mathscr{N}_i$. The normalization factor $((\#\mathscr{N}_i)\, M)^{-1}$ implies that $-1 \leq U(x,i) \leq 1$ for every $i \in \Lambda_N$ and every configuration x.

Finally, we introduce the Hamiltonian $\mathscr{H}(x)$, the opposite of the mean happiness of a given configuration x, defined as

$$\mathscr{H}(x) = -\frac{1}{N} \sum_{i \in \Lambda_N} U(x,i).$$

It is useful to consider the formal correspondence between particles trying to minimize the energy of the Hamiltonian, and

agents trying to maximize their utilities to define the system's dynamics. So, we fix some happiness threshold U_0, and we say that an agent located at i is unhappy if $U(i) < U_0$. Then, given some initial configuration x_0, we update the configuration by switching two unhappy agents, randomly selected, located at i, j with different signs (i.e., $x(i) \cdot x(j) = -1$).

The system will evolve until no unhappy agents can be found, or all unhappy agents have the same sign, so no partners are available for interchanging positions. We say that those final configurations are the stationary states of the model, and we can understand the stationary states as local minima of the Hamiltonian $\mathscr{H}(x)$ for this dynamic. The final configurations will show two or more clusters grouping agents of the same type. We are interested in the kind of minima obtained for different weights w. Simulations show that lower local minima of $\mathscr{H}$ are attained for slowly varying weights than for homogeneous weights over a region of the space.

RESULTS

We perform computational simulations of the model on one and two-dimensional lattices with periodic boundary conditions and for different weight functions. Each realization starts from a uniformly random distribution of agents' states. We show results obtained for the case where the populations of each type are of equal size and for a happiness threshold $U_0 = 0$. These results can be found in more detail in [1].

In Figure 1a we present final configurations for particular realizations of the one-dimensional case for two families of weight functions (w_k^1, w_k^2) that exhibit distinct number of oscillations over the 2500 sites. The black curve represents the weight and, at the top, there is the stationary state where agents in state 1 are blue and in state -1, red. In both cases, we see how a higher number of oscillations (that increases with k) relate to a larger amount of clusters in the final configuration. The segregation pattern shows small-scale clusters in this situation, similar to the classical Schelling model with constant weight.

We also implement simulations over a two-dimensional square lattice and weight functions that are constant over lattice columns. In the upper panels of Figure 1b, we show typical stationary states obtained for a square lattice of side 200, along with the corresponding weight function in greyscale. We see how few oscillations of the weight function (as in the left panel) generate larger scale segregation than a higher oscillating function (as in the right panel). In both cases, a more fragmented segregation pattern displays over the locally highly weighted sites. To quantify this effect, at the bottom of Figure 1b, we include plots of the mean number of cluster boundaries (i. e. the number of sign changes) within each lattice column, taken over 50 realizations with different initial distributions of the state of the agents. It is remarkable how the peaks of the sign changes per column coincide with the highest values of the weight.

In addition, we discuss some empirical examples. One of the variables that could give traces of the model's weight function is the territory's topography. The existence of hills, for instance, favors a propensity to appreciate the relations with the neighbors and to make communities [3].

In Figure 1c we present data on the topography, along with information about the ethnic distribution, for some regions of the Brazilian cities of Sao Paulo and Rio de Janeiro. In each city, we analyze an area of equal surface, defined from a square bounding box of 0.15 degrees on the sides of latitude and longitude. Rio de Janeiro shows more variability than Sao Paulo in the topography. To visualize information about the ethnic distribution of the population, we based on the IBGE (Brazilian Institute of Geography and Statistics) Census 2010 source [5]. The census provides geo-referenced data, at the census tract level, of the self-declared ethnicity of each Brazilian citizen (over five preset categories: White, Brown/Mixed, Black, Asian, and Indigenous (see [6]). In Figure 1c, we color each census tract according to its ethnic majority and observe different segregation patterns in each city. Although a quantitative analysis requires considering the density of the different ethnicities and the size of populations, a preliminary comparison of the ethnic segregation and topography maps could illustrate the idea of the high variation of a weighting function (associated with the topology of territory), smaller clusters arise.

DISCUSSION AND FURTHER WORK

Our model connects the traditional Schelling model where an agent wants to be surrounded by neighbors from the same group with a weight function associated with the relevance that the agent assigns to being surrounded by agents of the same type, which depends on each location. Hence, some places are relatively more important than others for establishing possible ties. Apart from the topography, another possible empirical interpretation of this weight function is the inverse of the price of land or real estate that each place represents. Another one is the inverse of the prestige of the area, the number of available resources, high provision of public goods, facilities, cultural and educational opportunities, transportation accessibility for the rest of the city, green spaces, and, from an aesthetics point of view, beauty areas. These general facilities are inverse to the weight functions of the places. And when these facilities are absent, the weight function takes a high value and the importance of ties to survive, too.

Under this rule, our model predicts two types of segregation patterns: a) higher spatially oscillations of the weighting function are associated with segregation levels similar to small patches or ghetto's formation, with clusters of smaller size; b) oscillations of the weighting function in the space corresponds to significantly segregated neighborhoods, with clusters of big size.

In what follows, we intend to apply the model to irregular networks that are more realistic and capture relevant geographic information about specific cities. Also, we plan to advance in the characterization and quantification of residential segregation based on census tract information, considering the distribution of socio-cultural groups of each census tract, not

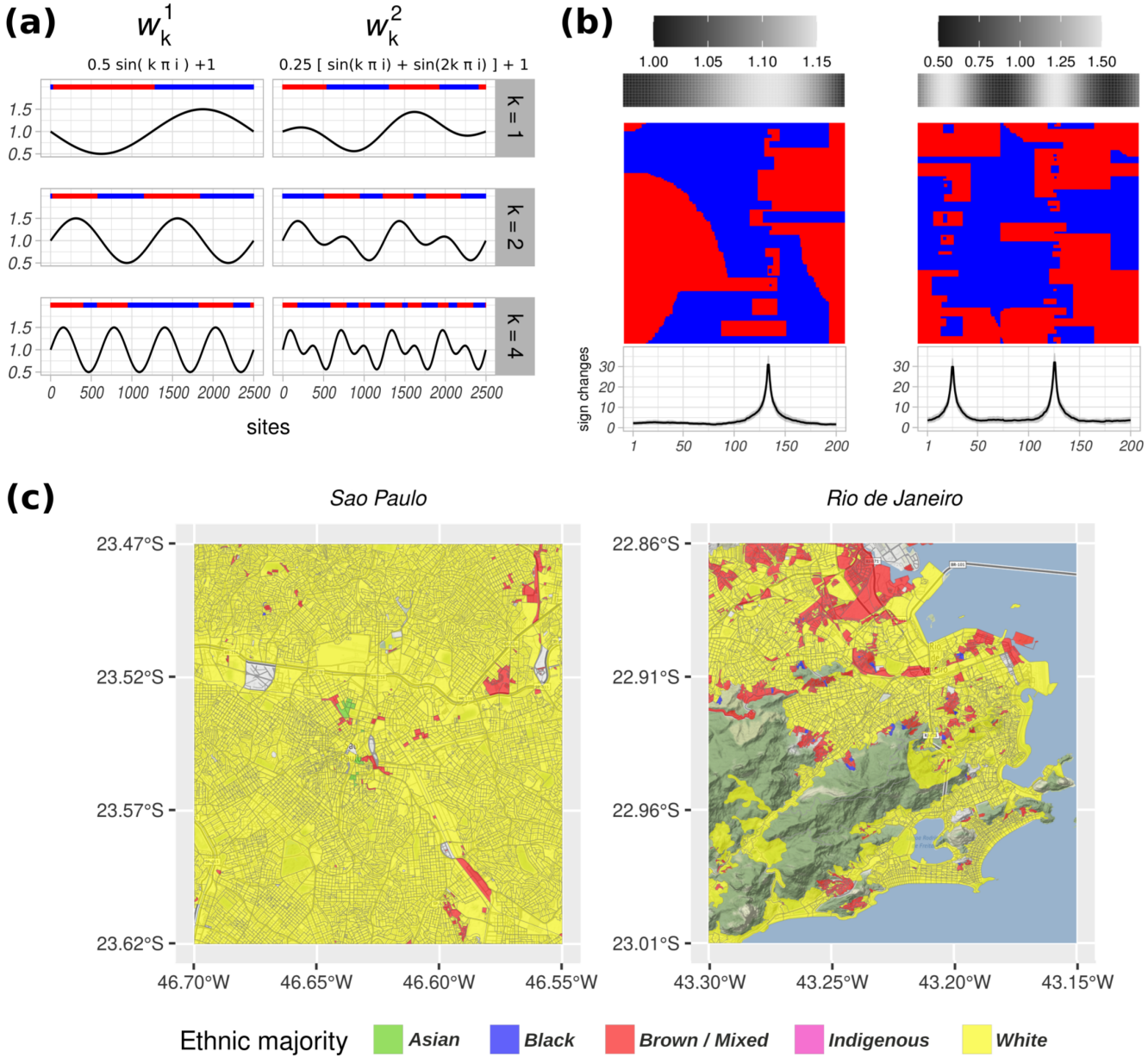


Fig. 1: (a) Stationary states on a one-dimensional lattice with $N = 2500$ nodes, happiness threshold $U_0 = 0$ and two families of oscillating weight functions (black curves). Top: final configuration, where agents in state 1 are blue and in state 1, red. (b) Stationary states on a squared lattice of side 200. The weight functions are at the top in greyscale; at the bottom, the mean number of sign changes within each column taken over 50 realizations with different initial distributions of agents states. (c) Topography maps and distribution of majority ethnicity according to 2010 Census for urban areas of Sao Paulo and Rio de Janeiro.

only individually but also concerning its neighboring tracts.

References

1. Arcon V., Caridi I., Pinasco J.P., Schiaffino P. (2022) Segregation patterns for non-homogeneous locations in Schelling's model (under evaluation).
2. Arcon V., Pinasco J.P., Caridi I.(2022) A Schelling-opinion model based on integration of opinion formation with residential segregation. Causes and Symptoms of Socio-Cultural Polarization, Springer.
3. dos Santos Oliveira N. (1996). Favelas and Ghettos: Race and Class in Rio de Janeiro and New York City. Latin American Perspectives, 23(4), 71-89.
4. Hatna, E., Benenson, I. (2012). The Schelling Model of Ethnic Residential Dynamics: Beyond the Integrated - Segregated Dichotomy of Patterns. Journal of Artificial Societies and Social Simulation.
5. Instituto Brasileiro de Geografia e Estadistica. http://https://censo2010.ibge.gov.br/
6. Patadata. Mapa interativo de distribuição racial no Brasil https://patadata.org/maparacial/en.html
7. Sabatini F. (2003). La segregación social del espacio en las ciudades de América Latina. Banco Interamericano de Desarrollo. Washington DC.
8. Schelling, T. C. (1969). Models of segregation. The American Economic Review, 488-493.
9. Schelling, T. C. (1971). Dynamic models of segregation. Journal of Mathematical Sociology, 1(2), 143-186.

Wealth distribution on a dynamic complex network

Gustavo L. Kohlrausch, Sebastián Gonçalves
Instituto de Física - Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brazil

Although an old problem, income and wealth disparities have increased substantially since the early XXI century [1, 2]. Despite the geographical, cultural, and historical differences, income distribution in different countries follows the same pattern [3], and understanding the mechanisms responsible for the origin and growth of inequalities is crucial to avoid it. Economic models based on statistical physics methods proposed in recent years [4] try to shed some light on this problem. In particular, agent-based models, where an agent is characterized by its wealth ω_i and savings fraction α_i. During the dynamics of the model, the agents exchange wealth, interacting in pairs following certain rules. One of the main strengths of this class of models is that different factors can be easily incorporated and studied, such as different rules for taxes and wealth exchanges [5].

On the other hand, a growing and promising field of study in economics is complex network theory [6]. In this approach, financial institutions or economic agents are the network's nodes, and relations among them are the edges. However, few contributions include complex networks into agent-based economic models [7, 8], and even in those, the network topology is fixed in time. Our goal is to provide a framework where a dynamic complex network is incorporated into an agent-based model, enabling a topological analysis of economic inequality phenomena. To study the disparities arising from economic transactions, we consider a conservative market, thus excluding the influences of processes such as the production of wealth and capital appreciation.

To create a dynamic network related to the wealth exchange process, we propose a model that considers that an agent's degree depends on its wealth. We justify this idea by considering that wealthier financial institutions can create more diversified investment portfolios. In the same way, wealthier companies or industries can reach a more significant number of investors or consumers. Thus, we propose a model which alternates between two dependent processes: the exchange of wealth between connected agents and the rewiring of the network connections. For the exchange of wealth we use the the yard-sale model, where the wealth exchanged between agents i and j is the minimum that the agents put at stake, $dw = \min[\alpha_i\omega_i, \alpha_j\omega_j]$. The wealth of agents i and j after the exchange is $w_i(t+1) = w_i(t) + dw$, and $w_j(t+1) = w_j(t) - dw$, so the total wealth is conserved. We assume a probability of the poorest agent winning the transaction given by

$$p_{i,j} = \frac{1}{2} + f \times \frac{|\omega_i(t) - \omega_j(t)|}{\omega_i(t) + \omega_j(t)}, \tag{1}$$

where f is the social protection factor, which varies from 0 to 1/2. The rewiring process starts by randomly selecting a pair i, j of agents, if this pair is disconnected the probability of creating a new connection follows

$$P_{i,j} = \frac{\omega_i(t) + \omega_j(t)}{\sum_l \omega_l(t)}, \tag{2}$$

where the sum in l is only on agents with at least one connection and $\omega_i(t)$ is the wealth of agent i at time t. If the pair is already connected, the link breaks with the complementary probability, $Q_{i,j} = 1 - P_{i,j}$. In order

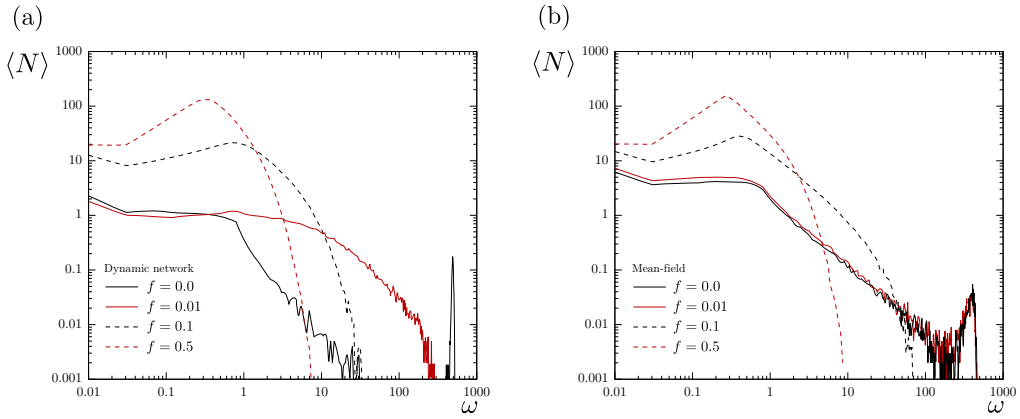


Figure 1: Non-accumulated distribution of wealth for different values of social protection in the (a) dynamic network and (b) mean-field model.

to all the agents being able to participate in this process, we randomly select $N/2$ pairs of agents to rewire their connections. After this process, we go back to the wealth exchange dynamics over the updated network structure. In this way, the system's evolution depends on wealth exchange and rewiring processes, which are dependent. The two process, wealth exchange between all the connected pairs, and the network rewiring of $N/2$ pairs, defines one Monte Carlo Step (MCS). We perform simulations for systems with size $N = 10^3$ over $t = 4 \cdot 10^4 MCS$. The results are averaged over 10^3 independent samples.

We obtain results for different values of f, analyzing economic and topological indicators, such as the Gini index, the distribution of wealth (Fig 1), assortativity (Fig 2), and the degree distribution. For the economic indicators, we compare our results with a mean-field model.

As shown in Fig 1, in the absence of social protection both models point out a condensation of wealth. Nevertheless, in the present model there is a gap in the wealth distribution, indicating a strong separation of classes, not present in the mean-field model results. The divergences between the two models are even more clear for $f = 0.01$, as the mean-field seems to be very robust to small increases of f, the dynamic network model presents a very distinct phenomenology from that of $f = 0.0$. As the social protection factor increases, the distributions of wealth for both models approach each other, yet the dynamic network model leads to smaller values of the Gini index. This results can be explained by the non-assortative behavior of the network as $f \to 0.5$, so the degree of the agents is not correlated with their neighbours (Fig 2).

In summary, our results showed that the dynamic network process has strong consequences in the agent based model, presenting a much richer phenomenology, especially in lower values of f. In our opinion, this simple model is able to reproduce distinct topological features which can be related to real societies, such as social stratification and the marginalization of the poorer agents.

Keywords: Econophysics, Wealth distribution, Economic networks, Agent-based models

Acknowledgments

This work was supported by the Brazilian funding agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

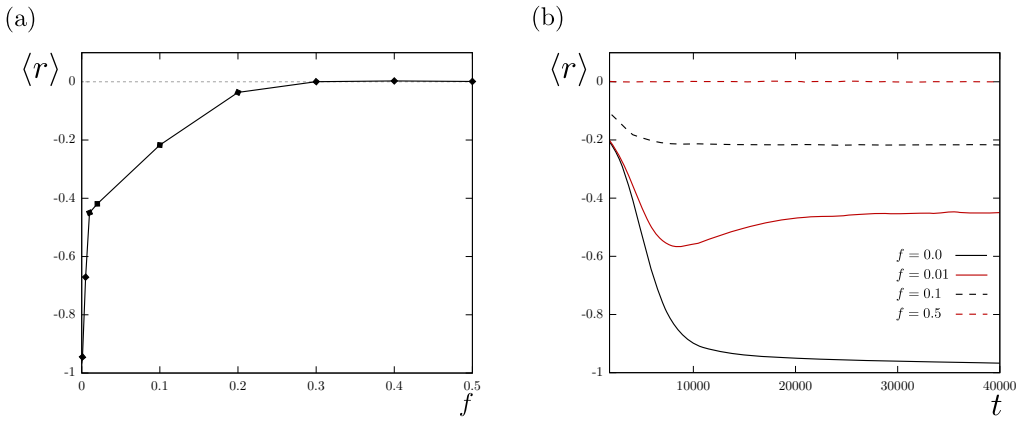


Figure 2: Assortativity of the network as a function of (a) social protection f and (b) time t for different values of f.

References

[1] L. Chancel, T. Piketty, Global Income Inequality, 1820–2020: the Persistence and Mutation of Extreme Inequality, Journal of the European Economic Association 19 (6) (2021) 3025–3062. doi:10.1093/jeea/jvab047.

[2] E. Saez, G. Zucman, Wealth inequality in the united states since 1913: Evidence from capitalized income tax data, The Quarterly Journal of Economics 131 (2) (2016) 519–578.

[3] B. K. Chakrabarti, A. Chakraborti, S. R. Chakravarty, A. Chatterjee, Econophysics of income and wealth distributions (2013).

[4] V. M. Yakovenko, J. B. Rosser Jr, Colloquium: Statistical mechanics of money, wealth, and income, Reviews of modern physics 81 (4) (2009) 1703.

[5] J. R. Iglesias, B.-H. F. Cardoso, S. Gonçalves, Inequality, a scourge of the xxi century, Communications in Nonlinear Science and Numerical Simulation 95 (2021) 105646.

[6] M. Bardoscia, P. Barucca, S. Battiston, F. Caccioli, G. Cimini, D. Garlaschelli, F. Saracco, T. Squartini, G. Caldarelli, The physics of financial networks, Nature Reviews Physics (2021) 1–18.

[7] T. Ma, J. G. Holden, R. Serota, Distribution of wealth in a network model of the economy, Physica A: Statistical Mechanics and its Applications 392 (10) (2013) 2434–2441.

[8] L. A. Braunstein, P. A. Macri, J. R. Iglesias, Study of a market model with conservative exchanges on complex networks, Physica A: Statistical Mechanics and its Applications 392 (8) (2013) 1788–1794.

Where the Landlords Are: Identification of Regions and Megaregions through Networks of Rental Ownership

Benjamin Preis
Massachusetts Institute of Technology, Department of Urban Studies and Planning
benpreis@mit.edu

Abstract

The US is home to more than 100 million renters, and approximately 11 million landlords, yet these two sides to the rental market are rarely studied in tandem. This study uses a multiscalar network-based approach to identify regions and interconnections between regions of rental markets. Many US cities require landlords to acquire a license for each rental property they own. Building on this administrative data of rental property and landlord location, I define rental property networks as a spatial bipartite network, where landlords are connected to their properties, both of which exist in physical space. First, I simplify this network by extracting its backbone, defining a rental market area. This rent-based definition of a region provides for a clear boundary of rental markets based on renter and landlord concentration and location, which is a substantial improvement on commuting-flow approaches. I then compare these regions to existing a null model, commuting zones, and against administrative boundaries, and evaluate how they capture actual rents and urban boundaries. Second, I define rental housing megaregions based on common institutional ownership across regions, where distinct regional rental housing markets may have large concentrations of the same landowner. These megaregions are then evaluated based on their physical distance and other connections, such as through migration. Researchers and policymakers have historically viewed rentals in markets as unconnected nodes. By identifying appropriate definitions of regions and megaregions through the creation and analysis of a rental ownership network, this research contributes to the literature on delineation and extent of rental market areas. It therefore provides a more robust foundation to understand rental market dynamics and the relationship between owner, renter, and property.

Keywords: urban networks, housing markets, spatial analysis, urban flows

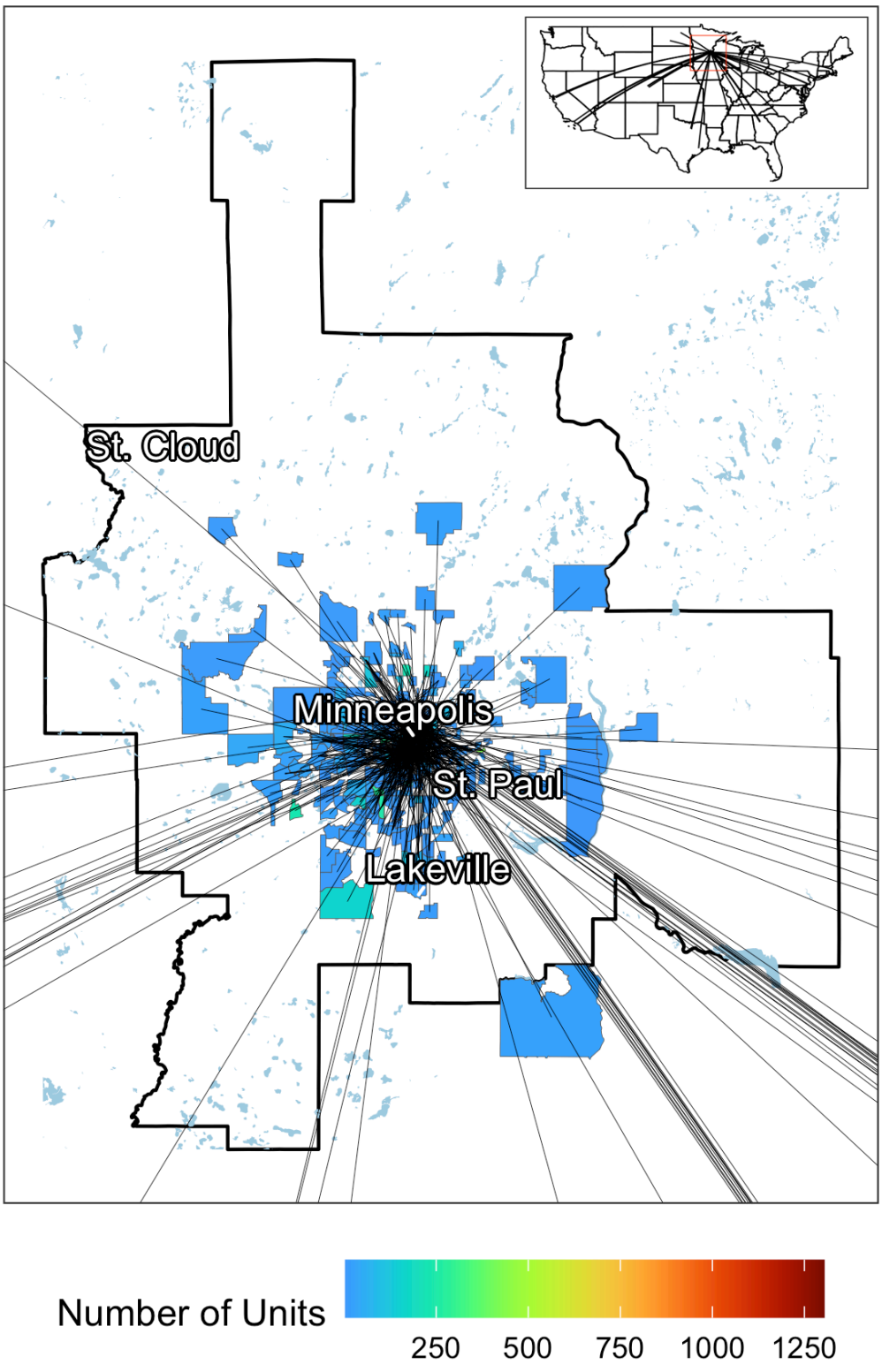


Figure 1: The Network Backbone of the Rental Ownership Market in Minneapolis, MN, USA. Extracted with the backbone package in R using a disparity filter. Overlaid on the Metropolitan Statistical Area definition with a black border. The national map in the top right hand corner shows how the rental property backbone extends throughout the entire US.

References:

Dash Nelson, Garrett, and Alasdair Rae. 2016. “An Economic Geography of the United States: From Commutes to Megaregions.” Edited by Joshua L Rosenbloom. *PLOS ONE* 11 (11): e0166083. https://doi.org/10.1371/journal.pone.0166083.

Duranton, Gilles. 2021. “Classifying Locations and Delineating Space: An Introduction.” *Journal of Urban Economics*, Delineation of Urban Areas, 125 (September): 103353. https://doi.org/10.1016/j.jue.2021.103353.

Khalife, Sammy, Jesse Read, and Michalis Vazirgiannis. 2021. “Structure and Influence in a Global Capital–Ownership Network.” *Applied Network Science* 6 (1): 16.

Neal, Zachary P. 2022. “Backbone: An R Package to Extract Network Backbones.” *PLOS ONE* 17 (5): e0269137. https://doi.org/10.1371/journal.pone.0269137.

Serrano, M. Ángeles, Marián Boguñá, and Alessandro Vespignani. 2009. “Extracting the Multiscale Backbone of Complex Weighted Networks.” *Proceedings of the National Academy of Sciences* 106 (16): 6483–88. https://doi.org/10.1073/pnas.0808904106.

Shelton, Taylor. 2018. “Rethinking the RECAP: Mapping the Relational Geographies of Concentrated Poverty and Affluence in Lexington, Kentucky.” *Urban Geography* 39 (7): 1070–91. https://doi.org/10.1080/02723638.2018.1433927.

Workers positional power

An input-output relations study.

Deborah Noguera
dnoguera@fahce.unlp.adu.ar

KEYWORDS: POSICIONALITY - WORKERS STRUCTURAL POWER - INPUT-OUTPUT NETWORK - PAGERANK CENTRALITY

1 Introduction: from posicionality to centrality

This paper analyzes the structure of inter-sectoral relations of Argentina's economy, via network analysis tools. The research question focuses on the notion of positionality and how it can explain the power relations between capital and labor. From our point of view, the position held by the different sectors in the economic system reflects the structural advantages of the actors involved in them.

The concept of posicionality has been addressed by different theoretical approaches from disciplines such as sociology, political economy and network theory. In this literature, position is used to highlight the multidimensionality of power relations and the structural importance of actors operating in a system. These contributions show that the strategic position of workers in an economic system gives them a "disruptive potential" to affect the normal functioning of the production process of key industries (Wright, 2000; Perrone et al., 1984). They also highlight the importance of divergent trajectories in sectoral profit rates (Marx, 1980; Botwinick, 2017) and the different levels of union organization and action (Barrera Insua and Marshall, 2019) as determinants of sectoral wages. Finally, actors position in an interconnected system as been widely approached through the analysis of its structural properties. In particular, the concept of centrality in network theory allows capturing the structural importance of actors in a system (Barabási, 2016). Therefore, centrality measures can be used as indicators of the structural power of actors through their position in the economic system.

In short, economic sectors hold a specific position in the production network, which give rise to a particular structure whose topological characteristics express the positional dimension of union bargaining power. The complex network theory concept that captures different aspects of a node's position is centrality (reflecting the actors structural importance); so we can operationalize the concept of structural power using these measures.

The aim of this work is: (1) to operationalize the concept of workers' positional/structural power through the analysis of the propierties of the production network in Argentina; and (2) to explore its link with the sectoral wages distribution.

2 Data and methodology

Wage negotiations take place at a sectoral level, so we are concerned with this scope of application: the degree of influence of workers actions in the negotiation with employers' organizations that represent firms belonging with different economic sectors (Barrera Insua and Marshall, 2019). Therefore, workers position will be determined by the network position held by the economic sector to which the firm where they work belongs. This information is provided by the Input-Output Table (IOT), obtained from the OECD database, which contains data for 45 sectors, according to the International Standard Industrial Classification system ISIC Rev.4[1], detailing the relationships between each of them at a national level. The data is annual, for the period 1998-2018 and for the Argentine case.

The IOT can be described as a network $G(V, E)$ directed and weighted, defined by te set of nodes V and the set of edges E. It is directed because IO systems represent bidirectional flows between economic sectors, *i.e.* each pair of nodes is connected by two links, one for each of the directions in which transactions may take place. It is also weighted because the links do not only represent the presence of a connection, but such connection has a specific magnitude.[2]

We propose the Weighted PageRank Index (WPR) (Brin and Page, 1998; Zhang et al., 2022) to approximate the structural power of workers. A node WPR score depends on: (1) it receives a large number of incoming edges, (2)

This is a global centrality measure that takes into account: (1) the neighbors position in the production network, (2) the number and weight of the incoming edges, (3) some node-specific quantifiable information attached to sector i and (4) a tuning parameter (θ) adjusting the relative importance of weights in the definition.

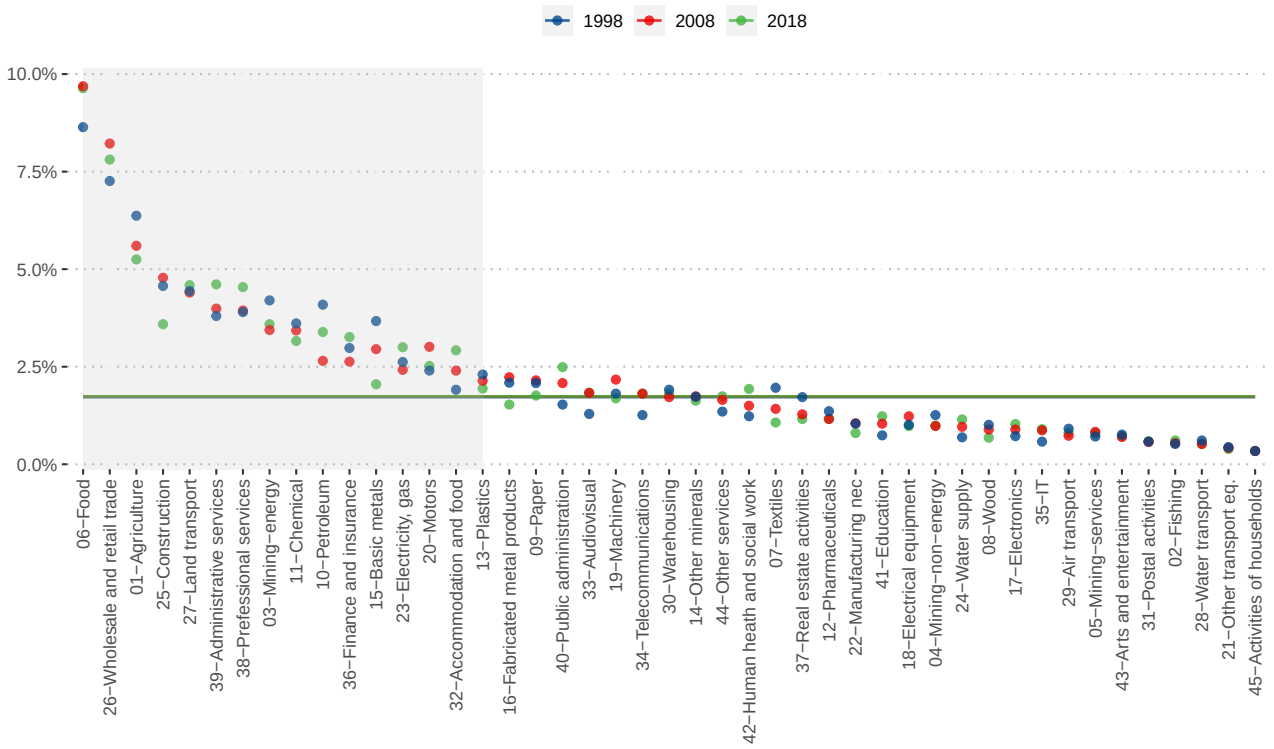
Formally, it is defined as:

$$\mathbf{wrank_i} = \alpha \sum_{j \in V} (\theta \frac{w_{ij}}{s_j^{out}} + (1-\theta) \frac{a_{ij}}{d_j^{out}}) \mathbf{wrank_j} + \frac{(1-\alpha)u_i}{\sum_{i \in V} u_i}$$

[1] The sectors can be consulted at the following link: https://www.oecd.org/sti/ind/input-outputtables.htm.

[2] Given the nature of IOT, the weighted adjacency matrix $W = (w_{ij})$ is a non-negative square matrix, where each element w_{ij} represents the volume of transactions directed from the node i to node j. The associated binary adjacency matrix $A = (a_{ij})$ is such that each element a_{ij} is equal to 1 when there is a link connecting node i with node j, and is equal to 0, otherwise.

Figure 1. Workers structural power, operationalized through WPR. Argentina, 1998,2008 y 2018



where $\theta \in [0,1]$[3]; d_j^{out}, s_j^{out} and $wrank_j$ are the out-degree, the out-strength and the WPR of node j, respectively; and u_i takes the non-uniform relative importance of the nodes into account. we include the share of each sector in formal employment in Argentina as relevant node-information to calculate its centrality[4].

3 Positional power and wage sectoral inequality

Interactions between sectors have remained relatively stable over time. However, we observe a slight growth in the last two decades, linked to the growing phenomenon of supply chain fragmentation that takes place on both global and local scale. However, regarding the top positions in the sectoral ranking, practically the same activities remain there, with some minor changes.

Figure 1 shows the ranking of structural power operationalized through the WPR. The ranking is led by the food industry, retail and wholesale trade, and the agricultural sector. Although it is a heterogeneous group, the three sectors are associated with global valorization processes, which is reflected in the fact that they are export-oriented industries.

We explore the link between structural power operationalized through the selected centrality measure and sectoral wages. When comparing WPR results and each sector position compared to the rest in terms of wages, we initially observe a positive correlation between both variables. Given the characteristics of employment and of the firms involved in the different sectors –among other factors affecting wage determination–, wage results of union intervention, reflected in sectoral wages, vary according to the more or less strategic position that each sector occupies in the economic structure and the consequent disruptive potential of union action.

We also explore the statistical distribution that these two variables follow, –namely, WPR centrality and sectoral wages– in order to assess whether they follow a power law[5], another distribution characterized byheavy-tails or neither. To do so, we follow the technique proposed by Clauset et al. (2009). The results indicate that there is not enough evidence to reject the null hypothesis that the sample comes from a power-law distribution[6].

It is important to notice that greater structural power does not necessarily mean higher wages or

[3]The value of θ can be chosen according to practical needs and actual interpretations. We are concerned with the volume of transactions, but also with the number of edges because it reflects the scope of the disruptive potential, so both degree and strength matter simultaneously. Therefore, we set $\theta = 0.5$

[4]For further details on the algorithm used, we refer the reader to Zhang et al. (2022).

[5]A power-law is a functional relationship between two quantities, which states that a relative change in one quantity results in a proportional relative change in the other, regardless of the initial size of those quantities. Mathematically, is expressed as $p(x) \propto x^{-\alpha}$, where α is the scaling parameter of the power-law distribution.

[6]As well as for a log-normal relationship. Up to know, we are unable to determine wheter power-law or log-normal better fits to out data. The results of the three stages provided by Clauset et al. (2009) are available upon request.

greater bargaining power overall. This is because: (1) structural power captures only a piece of the overall bargaining power, i.e. even in situations of high structural power, overall bargaining power may be low due to, for example, low associative power; y (2) the relationship between bargaining and wages is not deterministic but stochastic: high bargaining power means a high probability of success in the wage dispute and not directly higher wages.

Therefore, in order to get a more concrete idea regarding the relationship of the two variables, we regressed the logarithms of the two variables, including another relevant variables on sectoral wage determination as controls, based on Barrera Insua and Noguera (2021) analytical framework[7]. So, we estimate the exponent of the relationship

$$\mathbf{w} \sim WPR^{\alpha},$$

by regressing

$$\log(\mathbf{w}) = c + \alpha \log(WPR),$$

wehere $\mathbf{w}$ is the sectoral wage and WPR is the weighted PageRank centrality index. We found a statistically significant relationship with an exponent being on average around 1.9. Given that WPR centrality captures the relative workers structural power at the sectoral level, considering the national structure of production, the power-law between this variable and the sectoral wage implies that a relative change in the quantity of workers structural power may give rise to a proportional relative change in the quantity of sectoral wages, regardless the initial values of each variable.

Table 1. Sectoral wage model, results for WPR as a measure of structural power. Argentina, 2003-2018.

Dependent variable: mean sectoral wage (log)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log(PageRank)	**1.873**	**1.806**	**1.876**	**1.821**	**1.802**	**1.871**	**1.766**
	(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Controls	No	A1	A2	B	C	D	B,C,D
Fixed effects	Sí	Sí	Sí	Sí	Sí	Sí	Sí
N	945	945	945	945	945	945	945
R2	0.5382	0.5688	0.5409	0.5544	0.6201	0.5481	0.7164
adj. R2	0.538	0.569	0.541	0.554	0.620	0.548	0.716

Notes: p-values between brackets. Controls A=1-Total strength, 2-in/out strength. B=Sectoral profit rate, companies payment capacity. C= Unionization rate, conflict. D=Minimun wage.

[7]We include the following controls: the sectoral profit rate (Income Generation Account-CGI published by the National Institute of Statistics and Censuses-INDEC); the payment capacity of companies in each sector approximated by their average size (Ministry of Labor, Employment and Social Security-MTEySS); the unionization rate (National Survey of Workers on Conditions of Employment, Work, Health and Safety-ECETSS); the minimum wage (published in the Official Gazette); and the conflict, (variable built based on the information published by the MTEySS). For further details we refer to Barrera Insua and Noguera (2021).

4 Concluding remarks

In this paper we focus on studying the workers structural power and its link with wage inequality at the sectoral level. The main results can be summarized in the following two elements: (1) the distribution of positional power is asymmetric between sectors and that implies an asymmetric distribution of sectoral wages; (2) positional power is relevant to explain distributive conflict dynamics.

We consider that the paper's contribution is twofold. First, we propose an alternative way to measure the workers bargaining power, by operationalizing it through complex networks approach. On the other hand, we provide empirical evidence (at the national level) about the relationship between the workers structural power and the sectoral wage distribution.

References

Barabási, A. (2016). *Network Science*. Cambridge University Press, Cambridge.

Barrera Insua, F. and Marshall, A. (2019). Poder sindical en la negociación salarial. *Desarrollo Económico*, 59(228):251–270.

Barrera Insua, F. and Noguera, D. (2021). Determinantes salariales intersectoriales en la argentina: un modelo de análisis para las dinámicas desiguales del capital y el trabajo. In *15º Congreso Nacional de Estudios del Trabajo*. ASET.

Botwinick, H. (2017). *Persistent inequalities: wage disparity under capitalist competition*. Brill.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Marx, K. (1980). *Teorías sobre la plusvalía: tomo IV de El Capital*. Fondo de Cultura Económica.

Perrone, L., Wright, E. O., and Griffin, L. J. (1984). Positional power, strikes and wages. *American Sociological Review*, pages 412–426.

Wright, E. O. (2000). Working-class power, capitalist-class interests, and class compromise. *American Journal of Sociology*, 105(4):957–1002.

Zhang, P., Wang, T., and Yan, J. (2022). Pagerank centrality and algorithms for weighted, directed networks. *Physica A: Statistical Mechanics and its Applications*, 586:126438.