Proceedings Paper:
Escolano, Francisco, Lozano, Miguel Angel and Hancock, Edwin R. orcid.org/0000-0003-4496-2028 (2021) The Entropy of Graph Embeddings: A Proxy of Potential Mobility in Covid19 Outbreaks. In: Torsello, Andrea, Rossi, Luca, Pelillo, Marcello, Biggio, Battista and Robles-Kelly, Antonio, (eds.) Structural, Syntactic, and Statistical Pattern Recognition. Springer , Cham , pp. 195-204.

https://doi.org/10.1007/978-3-030-73973-7_19

The Entropy of Graph Embeddings: A proxy of Potential Mobility in Covid19 Outbreaks*

Francisco Escolano[1,2][0000−0003−3238−4021], Miguel Angel Lozano[1,2][0000−0002−4757−5587], and Edwin R. Hancock[3][0000−0003−4496−2028]

[1] University of Alicante, Spain
[2] COVID-19 Data Science Task Force
[3] University of York, UK

Abstract. In this paper, we predict the R_0 (reproductive number) of COVID-19 by computing the entropy of the mobility graph during the first peak of the pandemic. The study was performed by the *COVID-19 Data Science Task Force* at the Comunidad Valenciana (Spain) during 70 days. Since mobility graphs are naturally attributed, directed and become more and more disconnected as more and more non-pharmaceutical measures are implemented, we discarded spectral complexity measures and classical ones such as network efficiency. Alternatively, we turned our attention to embeddings resulting from random walks and their links with stochastic matrices. In our experiments, we show that this leads to a poweful tool for predicting the spread of the virus and to assess the effectiveness of the political interventions.

Keywords: Graph Embeddings · Graph Complexity · COVID-19.

1 Introduction

Motivation The outbreak of COVID-19 in Spain activated several research groups addressed to propose non-pharmacologic measures such as: (a) track the impact of global/local lockdowns and (b) model the impact of lockdowns in the progress of the infection. These are part of the objectives of the *COVID-19 Data Science Task Force*. This task force is formed by 20 interdisciplinar scientists of the main universities of the Comunidad Valenciana (CV). Our mission is to interpret, aggregate and make reports to policy makers. It has four areas: *mobile data analysis* (collect and geolocate anonymized cell phone data), *epidemiological models* (formulation and fitting of metapopulation models such as SIR or SEIR and agents-based ones, state of hospitals and ICUs), *predictive models* (hotspot detection, risk-priority maps, etc) and *citizen's science* (covid impact survey).

* BBVA Foundation, Banco de Santander and Spanish Government.

Outline of the paper The purpose of this paper is to show the link between the R_0 number of the SARS-CoV-2 and the complexity of the mobility graph. The R_0 (basic reproduction number) quantifies how many infectious cases are generated by a single one. Therefore for $R_0 > 1$ we have a spreading infection. The larger it is the more difficult is to control the infection. According to the Imperial College's Report on March 30, 2020[4], $R_0 \approx 5$ in Spain by March 9, 2020, when social distancing was implemented. This number was reduced to $R_0 \approx 1$ after the complete lockdown.

Global or partial lockdowns address the point of stopping the spread of the virus by reducing the mobility of people. This is why in the COVID-19 Data Science Task Force we started to work with anonymized mobility data. One of the objectives was to find a proxy of the spread of the virus by looking at mobility graphs. In principle, we wanted to enrich the prediction power of our stochastic epidemiological model (Section 2) where the effects of mobility where shadowed by the big numbers of the metapopulation model.

Then, we turn our attention to look at the topology of the mobility graphs. We followed two complementary strategies. One team studied the degree of fragmentation of the communities as the political measures were implemented. The second team (ours) interpreted the graph in a different way. More precisely, we looked at the stochastic matrices that encode the random walks potentially running on the network. Instead of dealing with a weighted digraph which is difficult to analyze by spectral means, we looked at the powers of the transition matrices. These matrices are the core to several recent embeddings such as *node2veec* [3], *Glove* [5] and *DeepWalk* [8] among others. Their unifying principle is to extract pairs of co-visted nodes and use these statistics (either by deep/shallow learning or SVD factorization) to find vectorial representations of the nodes. With these vectors at hand, one can use a vectorial complexity measure to find a correlation between R_0 and the topology of the graph. In Section 3 we show how the factorization of the expected co-ocurrence matrix leads to an informative embedding. This information is given by the rank of the co-occurrence matrix and this rank has deep implications in the complexity of several models of graphs. The *key idea* here is to relate the rank with the degrees of freedom of the topology. Summarizing, disconnected mobility graphs lead to low rank (i.e. redundant embeddings) since the random walks running on them are too constrained (they perform a few distinctive hitting patterns). However, more complex graphs are endowed with high-rank embeddings. Herein, the use of vectorial entropy estimator is a computational trick to bypass robust rank estimation.

In Section 4, we show our experiments with the SEIR model and the link between R_0 and the square of the vectorial entropy. In Section 5, we summarize our conclusions.

[4] `https://spiral.imperial.ac.uk:8443/handle/10044/1/77731`

2 Mobility in Epidemiological Models

We use cell-phone geolocation data[5] to track the spread of the SARS-CoV-2 within the Comunidad Valenciana (CV) in Spain. We build mobility networks to map 6.8 million people of 324 census block groups (CBGs) between March 15 and May 23, 2020. Each GBG has at least $5,000$ people.

Stochastic SEIR We overlay a metapopulation SEIR model in order to track the infection trajectories, predict the R_0 number and monitor the epidemiological status of the 24 Health Departments of the CV. Each CBG maintains four sub-population: susceptible (S), exposed (E), infectious (I), and removed (R). The differential equations governing these sub-populations are:

$$\begin{aligned} S_t &= \frac{X_t}{n} \\ E_t &= S_{t-1} - \frac{X_t}{n} + \frac{Y_t}{n} \\ I_t &= E_{t-1} - \frac{X_t}{n} + \frac{Z_t}{n} \\ R_t &= R_{t-1} + I_{t-1} - \frac{Z_t}{n} \end{aligned} \tag{1}$$

where X, Y and Z are binomial distributions:

$$X_t \sim \mathbb{B}(nS_{t-1}, e^{-\beta I_{t-1}}),\ Y_t \sim \mathbb{B}(nE_{t-1}, e^{-\sigma}),\ Z_t \sim \mathbb{B}(nI_{t-1}, e^{-\gamma})\ . \tag{2}$$

and: n is the population, $\sigma = 1/5.1$, $\gamma = 1/12$ and $\beta = R_0\gamma$. The most important parameter is R_0, the reproduction rate (or reproduction number), which indicates the expected number of infectious cases generated by one case. In order to incorporate mobility to the model, we have to consider that the population is divided into N CBGs. As a result we have the conservation rule: $S_t + E_t + I_t + R_t = S_0 + E_0 + I_0 + R_0 = 1$ for all t and

$$S_t = \sum_{i=1}^{N} S_t^{(i)},\ E_t = \sum_{i=1}^{N} E_t^{(i)},\ I_t = \sum_{i=1}^{N} I_t^{(i)},\ R_t = \sum_{i=1}^{N} R_t^{(i)}\ . \tag{3}$$

The above decomposition applies also for the binomial distributions and we have:

$$X_t = \sum_{i=1}^{N} X_t^{(i\to j)}\ , Y_t = \sum_{i=1}^{N} Y_t^{(i\to j)}\ , Z_t = \sum_{i=1}^{N} X_t^{(i\to j)}\ , \tag{4}$$

where the superscript $(i \to j)$ denotes how many movers in the corresponding state for the $i-$th CBG move to the $i-$th CBG.

Using this model we predict the infectious cases for the whole CV (see Fig. 1).

[5] Provided the INE (National Institute of Statistics) due to an agreement between the Spanish Government and the main phone operators. These data are anonymized and register displacements between INE-GBGs

Mobility Graphs and Radiation Model The daily movers between CBGs create a *mobility graph* $G_t = (V, \mathcal{E}_t, \mathcal{W}_t)$ where $|V| = N$, and an edge $\epsilon_{ij} \in \mathcal{E}_t$ exists when $\mathcal{W}_t(i,j) = \frac{M_t(i,j)}{\sum_k M_t(i,k)} > 0$, where $M_t(i,j)$ are the movers from node i to node j at time t.

However, as we have only 324 CGBs we must introduce as much information as possible in order to model mobility fluxes properly. Therefore we use the so called *radiation model* [11] which takes into account the populations of the commuting CGBs as well as the populations of the CGBs in between. In this model $\mathcal{W}_t(i,j)$ is multiplied by

$$T_{ij} = \frac{N_i N_j}{(N_i + S_{ij})(N_i + N_j + S_{ij})} \tag{5}$$

where: N_i and N_j are the populations of CGBs i and j, and S_{ij} is the number of people in a circle centered at i with radius r_{ij}. This model is parameter-free (wrt to others such as the gravitational one) and it predicts better thr probability of observing a flux given the known distribution of the populations (origin, destination, in-between).

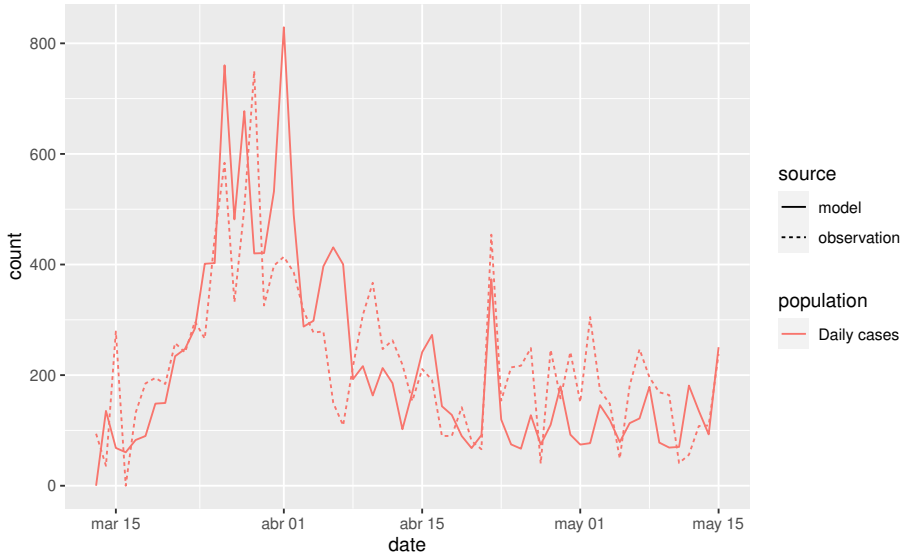

Fig. 1. SEIR model: Observed vs predicted cases

3 Entropy of Mobility Graphs

3.1 Embeddings and Random Walks

.

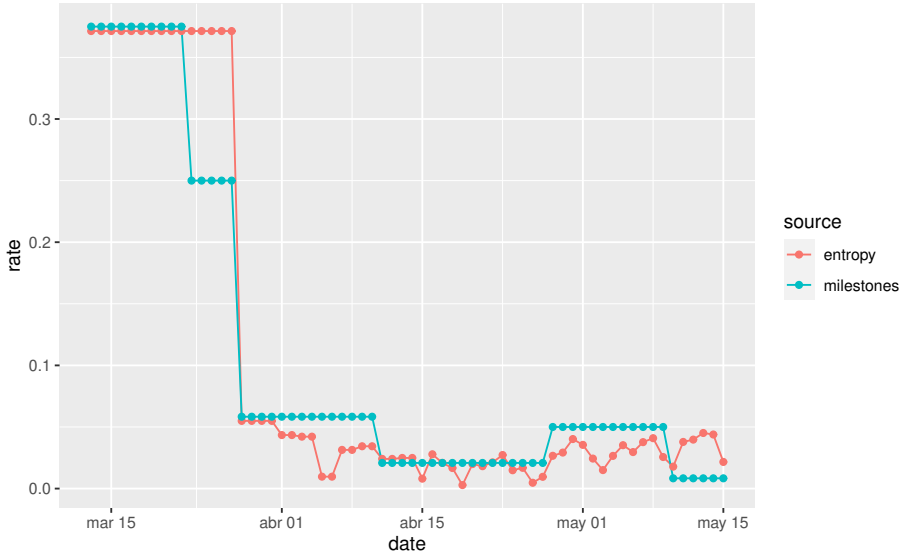

Fig. 2. Correlation between R_0 and Graph Entropy.

Expected Random Walks Following the standard factorization approach for network embedding [6], the latent representations for nodes of $G = (V, E)$ are obtained from the SVD of $\hat{\mathbf{M}}_G = \log(\max(\mathbf{M}_G, 1))$, where $\mathbf{M}_G$ is the Pointwise Mutual Information (PMI) matrix. More recently, Qiu et al. [9] show that $\mathbf{M}_G$ can be posed in the following terms

$$\mathbf{M}_G = \frac{vol(G)}{b}\mathbf{S}_G,\ \mathbf{S}_G = \left(\frac{1}{T}\sum_{r=1}^{T}\mathbf{P}_G^r\right)\mathbf{D}_G^{-1}, \tag{6}$$

where $\mathbf{P}_G = \mathbf{D}_G^{-1}\mathbf{W}_G$ is the transition matrix of G. This emphasizes the role of the random walks (RWs) in the resulting embedding. For instance, following [2] , $\mathbf{S}_G$ can be seen as the expectation of the the *co-ocurrence matrix* $\mathbf{O}_G \in \mathbb{R}^{N\times N}$ where the $\mathbf{O}_{G_{ij}}$ entry contains the number of times nodes i and j are co-visited within a context distance T, i.e. the number of times that a random walk starting an any node hits both i and j at most T steps. The hyperparameter T is called the *window size* and it controls the extent of a nodal context. Thus, for a fixed T Abu-El-Haija et al. define:

$$\mathbb{E}[\mathbf{O}_G; T] = \left(\sum_{r=1}^{T} Pr(c \geq r)\mathbf{P}_G^r\right)\mathbf{P}_G^0, \tag{7}$$

where $Pr(c \geq r)$ is the probability that node c belonging to the context of any *anchor node* is reached after r steps. Consequently, $\mathbf{S}_G = \mathbb{E}[\mathbf{O}; T]$ results from assuming: (a) $Pr(c \geq r) = \frac{1}{T}$, and (b) $\mathbf{P}_G^0 = \mathbf{D}_G^{-1}$. These simplifying assumptions lead respectively to: (a) Nodes within a context are chosen uniformly and

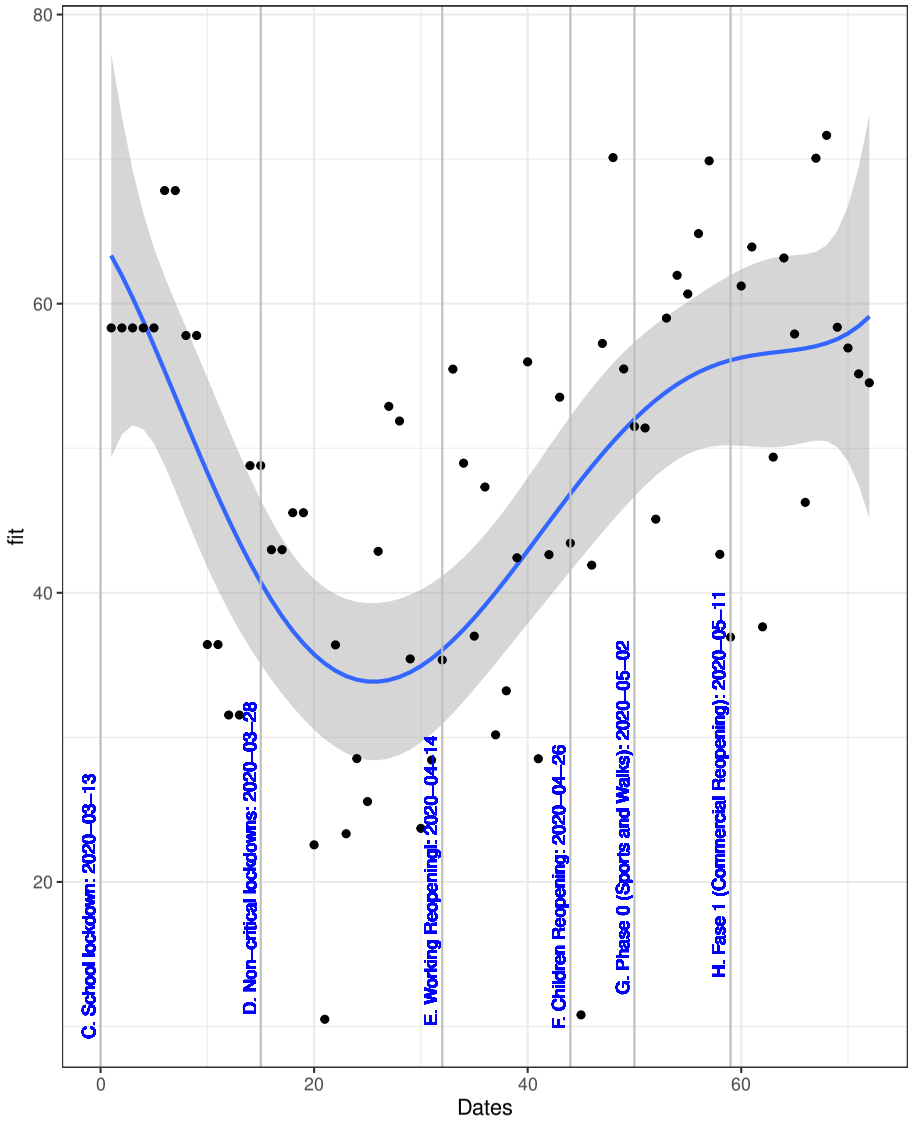

Fig. 3. Fitted (blue line) v Real (black dots) graph entropy

independently of how deep are the random walks, and (b) The probability that a random walk starts at a given node i are inversely proportional to its degree $\mathbf{d}_i$. Then, looking at $\mathbf{S}_G$ we can interpret $\mathbb{E}[\mathbf{O}_G; T]$ in light of the powers $\mathbf{P}_G^r$ of the transition matrix. More precisely, since the embedding relies on the SVD of $\mathbf{S}_G$, we herein *propose to relate the entropy of G with the rank of $\mathbf{S}_G$*.

Rank and Complexity We commence by exploring random graphs $G(N, \rho, \mathbf{Q})$ with a community structure. The rank of $\mathbb{E}(\mathbf{P}_G)$ relies on that of $\mathbb{E}(\mathbf{A}_G)$ which in turn is upper bounded by the rank of the $K \times K$ *communicability matrix* $\mathbf{Q}$, where K is the number of communities. The parameter ρ is a discrete probability distribution, i.e. $\sum_{k=1}^{K} \rho_k = 1$, at it induces a node generator τ, where a given node belongs to the $k-$th community with probability ρ_k. In addition, there is an edge between two nodes i and j there is an edge between them with probability $\mathbf{Q}_{\tau(i),\tau(j)}$. This is the so called *Stochastic Block Model* (SBM) [1] and its used for community detection. We consider two cases:

a) If $\mathbf{Q}$ is *symmetric* we have $\text{rank}(\mathbb{E}(\mathbf{A}_G)) \leq K$ because the adjacency matrices generated via SBMs have a block structure. One effective way of increasing the latter upper bound is to minimize the entropy of ρ. This leads to (almost) full-connected (complete) dis-assortative graphs with $\text{rank}(\mathbb{E}(\mathbf{A}_G)) \approx N$. This includes *ego-nets*, that is networks that code social circles with a large overlap between them such as the Facebook net [7].
b) If $\mathbf{Q}$ is *not symmetric* it is the degree of asymmetry what determines whether $\text{rank}(\mathbb{E}(\mathbf{A}_G)) \approx N$ (the larger the better) or not, independently of the entropy of ρ. This includes *citation networks* such as Cora [10] [6], CiteSeer for Document Classification [10] [7], and Wiki [8]).

Summarizing, dense strictly directed graphs achieve the largest ranks for $\mathbf{A}_G$ and consequently for $\mathbf{P}_G$. This is the case of the mobility networks studied in this paper. This is key, because usually $\text{rank}(\mathbf{S}_G) \leq \text{rank}(\mathbf{P}_G)$ (matrix powers and matrix addition do not preserve, in general, the rank).

The above facts lead us to interpret $\text{rank}(\mathbf{S}_G) \equiv \text{rank}(\mathbb{E}[\mathbf{O}_G; T])$ as a proxy of the *complexity of G*:

a) *Low complexity.* The rank determines the number of independent subspaces of the expected co-ocurrence matrix. Thus, $\text{rank}(\mathbb{E}[\mathbf{O}_G; T]) = p$ with $p \ll N$ indicates that the hitting patterns of the RWs are highly redundant, i.e. they collapse in a small number of p clusters, jointly visiting the same nodes in each cluster. Such a redundancy reveals that the random walks are constrained to hit the same subset of nodes independently of how far are

[6] Citation network containing 2708 scientific publications with 5278 links between them. Each publication is classified into one of 7 categories.
[7] Citation network containing 3312 scientific publications with 4676 links between them. Each publication is classified into one of 6 categories.
[8] Contains a network of 2405 web pages with 17981 links between them. Each page is classified into one of 19 categories. `https://github.com/thunlp/MMDW/`

them from their anchors. As a result, low rank means also low transport efficiency (and also low graph complexity) without relying on the inverse lengths of the shortest paths.

b) *High complexity.* When p is large (ideally $p \approx$ N) the co-occurrence patterns are linearly independent because the RWs are less constrained. The absence of bottlenecks favors transport efficiency due to the higher complexity of the graph.

Rank and Entropy Since rank estimation can be shadowed by numerical errors [12], p typically over-estimates the number of real co-ocurrence clusters. We herein address this problem by estimating the entropy of the embedding. Therefore,

(1) *Embedding.* Let $\mathbf{E}_G = \mathbf{U}_d\sqrt{\Sigma_d}$ the embedding matrix given by the rank$-d$ approximation of $\log(\max(\mathbf{M}_G, 1)) \approx \mathbf{U}_d\Sigma_d\mathbf{V}_d^T$ where $\mathbf{M}_G$ relies on $\mathbf{S}_G$ (see Eq. 6) i.e. on $\mathbb{E}[\mathbf{O}_G; T]$. Then, the $i-$th row e_i of $\mathbf{E}_G$ is the $d-$dimensional embedding of the $i-th$ node of G.

(2) *Bypass Entropy.* Given N $d-$dimensional points, their $\alpha-$Rényi entropy H_α, with $\alpha > 1$ is consistently estimated by the following functional [4]:

$$\hat{H_\alpha} = \frac{1}{1-\alpha}\log\frac{L_p(F)}{N^{1-p/d}}\ , \tag{8}$$

where: $p = d(1-\alpha)$, $F = (\mathcal{V}, \mathcal{E})$ is a $k-$nn graph whose vertices are the $\mathbf{x}_i = \mathbf{e}_i$ the embedding vectors, the edges $\mathcal{E}_{ij}$ are provided by the k-nn rule. Therefore we have

$$L_p(F) = \sum_{\mathcal{E}_{ij}} ||\mathbf{x}_i - \mathbf{x}_j||^2 \tag{9}$$

and γ is a normalization constant that can be estimated by generating a large sample of points in $[0,1]^d$ and running the estimator in its $k-$nn graph.
Thus, given the embedding, the bypass entropy returns its entropy. It is desirable to choose $\alpha \approx 1$ (close to the Shannon entropy). In this work we set: $\alpha = 0.99$, $p = 2$ and $k = 4$. The embedding vectors are also normalized before computing the entropy.

4 Experiments

Setting: Mobility Flows Following the INE protocol[9], for each CBG, a cell phone operator provides the number of terminals that are going to be considered as *living poputation*: owners of cell phone who spend most of the time at that CBG from 00:00-6:00am. This is the source CGB, whereas the destination CGB is the most visited CGB from 10:00am-16:00pm if the owner is there at least for 2 hours. The operators (Telefónica, Orange, Vodafone) report the *number of movements* to the destination CGB if there are at least $10-15$ movements (depending on the operator).

[9] National Institute of Statistics: `https://www.ine.es/covid/covid_movilidad.htm` (Techical Project).

Predicting R_0 Our stochastic SEIR model fits well the infection cases (see Fig. 1). However, in absence of any other information the R_0 parameter must be adjusted daily and specially after imposing a non-pharmaceutical measure (social distancing, lockdown). In practice, R_0 can be seen as control parameter that encodes all the measures implemented to slow down the propagation of the virus. However, as most of these measures are somewhat related to mobility we conjectured that there should be a mathematical relationship between the entropy of the embedding (which reflects the degrees-of-freedom of mobility) and R_0. After experimentation, we found that

$$R_0 \propto (H_\alpha)^2 \, . \tag{10}$$

as we show in Fig. 2, where we plot R_0 and the above estimator at several milestones during the lockdown. We give more details of the milestones and the curve fitting of the entropy in Fig. 3.

With this mathematical tool at hand we could monitor not only the global evolution of the CV but also the evolution of each of its 24 Health Departments during the first peak of the pandemic.

5 Conclusions

In this paper, we have proposed and successfully tested a proxy of the R_0 number via the complexity of the mobility graph. Such a complexity measure is closely related to the rank of state-of-the-art matrices which encode co-visiting statistics. The key idea is to relate the complexity of a graph with the degrees-of-freedom of the random walks running on it. It these random walks are constrained then the graph is simple (e.g. fragmented as it the COVID-19 mobility graph after political interventions); otherwise the graph is complex. We bypass the robust estimation of the rank by using a vectorial entropy estimator.

Future work includes the validation of this model in larger graphs as well as exploring the links between the proposed complexity and other alternatives. The underlying idea is to make this proxy much closer to an early warning system.

Acknowledgements The *COVID-19 Data Science Task Force* is founded by two Spanish Banks: BBVA and Santander, which have opened competitive calls for funding. BBVA funds the project: AYUDAS FUNDACIÓN BBVA A EQUIPOS DE INVESTIGACIÓN CIENTÍFICA SARS-CoV-2 y COVID-19. Santander funds the FONDOS SUPERA. In addition, M.A. Lozano and Francisco Escolano are funded by the project SPRIT (INFORMATION THEORY FOR STRUCTURAL PATTERN RECOGNITION, code RTI2018-096223-B-I00)of the Spanish Government. The COVID-19 Task Force is also grateful to the INE (National Institute of Statistics) who provides the data and the Generalitat Valenciana for its support.

References

1. Abbe, E.: Community detection and stochastic block models: Recent developments. J. Mach. Learn. Res. **18**(1), 6446–6531 (Jan 2017)
2. Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., Alemi, A.A.: Watch your step: Learning node embeddings via graph attention. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31, pp. 9180–9190. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper/2018/file/8a94ecfa54dcb88a2fa993bfa6388f9e-Paper.pdf
3. A.Grover, Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 855–864 (2016)
4. Escolano, F., Suau, P., Bonev, B.: Information Theory in Computer Vision and Pattern Recognition. Springer Publishing Company, Incorporated, 1st edn. (2009)
5. J.Pennington, Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 1532–1543 (2014)
6. Levy, O., Goldberg, Y.: Neural word embedding as implicit matrix factorization. In: Advances in Neural Information Processing Systems 27. pp. 2177–2185 (2014)
7. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. In: Bartlett, P.L., Pereira, F., Burges, C.C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 548–556 (2012), http://papers.nips.cc/paper/4532-learning-to-discover-social-circles-in-ego-networks
8. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014. pp. 701–710 (2014). https://doi.org/10.1145/2623330.2623732
9. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., Tang, J.: Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 459–467. WSDM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159706, http://doi.acm.org/10.1145/3159652.3159706
10. Sen, P., Namata, G., Bilgic, M., L.Getoor, Gallagher, B., T.Eliassi-Rad: Collective classification in network data. AI Magazine **29**(3), 93–106 (2008), http://www.aaai.org/ojs/index.php/aimagazine/article/view/2157
11. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012). https://doi.org/10.1038/nature10856, https://doi.org/10.1038/nature10856
12. Ubaru, S., Saad, Y.: Fast methods for estimating the numerical rank of large matrices. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. p. 468–477. ICML'16, JMLR.org (2016)