IMT School for Advanced Studies, Lucca
Lucca, Italy

The Complexity of Heterogeneity in Real-World Networks

PhD Program in
Institutions, Markets and Technologies - Systems Science
Track in Economics, Networks and Business Analytics

XXXIII Cycle

By

Matteo Serafino

July 30th, 2021

The dissertation of Matteo Serafino is approved.

PhD Program Coordinator: Prof. Pietro Pietrini, IMT School for Advanced Studies Lucca

Advisor: Dr. Tommaso Gili, IMT, School for advanced Studies Lucca

Co-Advisor: Prof. Guido Caldarelli, Department of Molecular Sciences and Nanosystems, Ca' Foscari University of Venice, 30172 Venice, Italy

Co-Advisor: Dr. Giulio Cimini, Department of Physics, University of Rome Tor Vergata, 00133 Rome, Italy

The dissertation of Matteo Serafino has been reviewed by:

Prof. Yamir Moreno, Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Zaragoza, Spain

Dr. Paolo Barucca, University College London, Department of Computer Science, London, UK

IMT School for Advanced Studies Lucca
July 30th, 2021

Alla nostra nuvola di Fantozzi...

Contents

List of Figures

List of Tables

Acknowledgements

My first acknowledgment goes to Guido Caldarelli. His wisdom guided me through the ups and downs of the Ph.D. His simplicity made everything simpler than it was. His "what do you want to do?" or "where do you want to go?" highlights that we can be whoever we want and where we want, but not without effort.

Giulio Cimini. He was with me through my all path, from the master thesis until today. Despite my mistakes and misunderstandings, he never gave up. Even when everything seems to be falling apart, he kept sustaining me. He taught me that patience is another ingredient, along with commitment.

Tommaso Gili. He is an example to follow both in terms of academic and personal life. Whenever I was losing my self-confidence, he was there to support me. During these three years, he has been my light-house. He taught me that spontaneity is what makes each recipe personal.

Hernan H. Makse. By working with him, I learned you never get tired to do what you like. I learned that fun and work do not necessarily need be two separate things.

The Network unit. Although I do not dedicate few lines to each of them, all the unit's components helped me during the Ph.D. I'm grateful to all of them for the exciting and insightful discussion.

But that's not all. This work would not be possible without the influence and support of my colleagues, friends, and family. Thanking all those who have marked my path would require tons of pages and shift the focus from the action to the

subject. The truth is that these people shaped my way of being, how I think, the way I approach people, and the way I do research. These people lit up my darkest days.

If there is something I learned by studying Complexity, it is that a mixture of elements behaves differently from the single entities. In such a scenario, without all these people, I may not have reached this point. From my perspective, each of these persons is a co-author of this work. Only bureaucracy kept me from entering all their names.

Vita

April 12, 1991 Born, Tricase, Italy

2014 **B.SC in Physics**
Thesis: "The Kronig-Penney model"
Final mark: 91/110
Sapienza University of Rome, Rome, Italy

2017 **M.Sc. in Physics**
Thesis: "Stochastic methods and network theory for the study andthe validation of financial time series"
Final mark: 110/110 cum laude
Sapienza University of Rome, Rome, Italy

Publications

1. M. Bruno, S. F. Sousa, F. Gursoy, M. Serafino, F. Vianello, A. Vranić and M. Boguñá. "Community Detection in the Hyperbolic Space.", preprint arXiv:1906.09082, 2019.
2. M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar and G. Caldarelli."True scale-free networks hidden by finite size effects". Proceedings of the National Academy of Sciences, vol 118 n° 2, 2021.
3. M. Serafino, H. S. Monteiro, S. Luo, S. DS. Reis, C. Igual, A. Lima, M. Travizano, J. Jr. Andrade and H. A. Makse."Superspreading k-cores at the center of COVID-19 pandemic persistence," under peer review at PLOS Computational Biology, preprint: arXiv:2103.08685, 2021.
4. Z. Zhenkun, M. Serafino, L. Cohan, G. Caldarelli and H. A. Makse."Why polls fail to predict elections," under peer review a European Physical Journal, preprint arXiv:2101.11389, 2021.
5. S. Feldman, J. Flamino, A. Galeazzi, B. Cross, Z. Zhenkun, M. Serafino, A. Bovet, H. A. Makse, and B. K. Szymanski. "Twitter Influencers and Increased Polarization during two U.S. Presidential Elections," under peer review a Proceedings of the National Academy of Sciences, 2021.

Presentations

1. M. Serafino, '"Inferring the political orientations of Twitter accounts from their connections: the contagion of ideas," Toffee Workshop #1, Lucca, Italy, 2019.
2. M. Serafino, "Scale free networks revealed from finite size effects," Conference on NetSci-X Tokyo, Japan, 2020.
3. M. Serafino, "Superspreading k-cores at the center of COVID-19 pandemic persistence," Conference on Complex Systems, Rome, Italy, 2020.
4. M. Serafino, "Inferring the political orientations of Twitter accounts from their connections: the contagion of ideas," (poster presentation) Conference on Complex Systems, Rome, Italy, 2020.
5. M. Serafino, "Superspreading k-cores at the center of COVID-19 pandemic persistence," APS March Meeting, Nashville, United State, 2021.

Abstract

Many real-world phenomena are characterized by complex structures. Modeling and detecting these architectures is of paramount importance to understand the dynamics of the considered systems and to consciously intervene on them. The COVID-19 pandemic is, in this sense, a telling example. One common feature of these complex structures is the heterogeneity of the node connectivity of the underlying network. This heterogeneity is one of the main culprits of the ongoing pandemic.

In this thesis, we introduce a method capable of uncovering complex networks' heterogeneous structures when finite-size effects hide the latter. Larger heterogeneity in the network structure leads to a smaller epidemic threshold. It is not a coincidence that policymakers worldwide are trying to reduce this heterogeneity (employing lockdowns and other less restrictive measurements) to stop the spread of the virus. We show that a macro-quarantine followed by a micro-quarantine can help in this direction. For this scope, we introduce an algorithm that is able to track super-spreaders. Notably, the same algorithm can be used to define an optimized strategy for vaccinations. Similarly to a virus, information continuously spreads on social networks. By combining Machine Learning and network theory techniques, we develop an algorithm able to discover what "social users" think about a particular topic—a sort of social listener, an AI alternative to traditional polls.

Chapter 1

Introduction

In the last three decades, we have witnessed a profound change in our world. The difference was so drastic that our ancestors would not recognize it as their planet. The main actors that drove this change are **technological progress** and its **availability for the public**. Let us consider one of the prominent inventions of the last century: the mobile phone. In 1984, the first-ever genuinely portable mobile phone hits the shelves, the Motorola DynaTAC 8000X (Despite the Federal Communications Commission's approval, on the 21st of September 1983). The mobile weighs 793 grams and could memorize up to 30 phone numbers. The phone took 10 hours to charge for a maximum of 60 minutes of calls. Its price was around four thousand dollars (about ten thousand dollars of today [1]). Although the technological change was there, the mobile phone was still a business tool because of its high price. When the mobile phone became more affordable, it led to a revolution in the way we communicate [2].

The mobile phone was the forerunner of a dynamic world. Quickly, the possibility to reach someone without stopping your activity became a need. At a certain point, the necessity of the mobile phone requested for a more performant battery. Then, to a lighter mobile phone. Later, to a mobile phone with a color display. Thanks to the Internet, mobile phones started to play a central role in daily life rapidly. A phone-call turned into a video-call. Finally, everyone could reach everyone, no mat-

ter where she or he was. All one needs is a phone number or a profile name (in the modern social networks). As a result, the words became hyper-connected. Mobile phones are just an example of the progress' speediness we are witnessing. We are filling up the words with technology and devices meant to make our life easier. Services designed to help people to carry out daily activities in new ways. We redesigned our world. In such a way that, every day, countless connections may take place. Physical as it is the flow of people and goods from one country to another one; virtual as the information traveling on the Internet on the world wide web and the services on it; financial with respect to the architecture of our financial systems with all the loans and credit amongst banks and firms [3].

Of course, "There is no such thing as a free lunch." Complexity is the quantity, whose increase gives a direction to evolution [4]. Complexity is the price that we are paying for evolution. It arises whenever the number of "actors" in the system becomes very large. When the law capable of explaining the behavior of the single ingredient, it fails in explaining their aggregations' behavior, confirming that "more is different" [5].

Although it is the first time that complexity "walks freely in our world," it was always with us, working, silently and efficiently, in the back-end. In 1872, to describe gas's behavior, composed of millions and millions of particles (N=10^{25}), Ludwig Boltzmann introduced the equation that carries his name. Gas, something present in our everyday life, is a manifestation of complexity. Statistical physicists, which deal with many elements, were the ones who first realized it.

At this point, a question arises: Why just now? Why did complexity not continue to work in the back-end? The answer is simple: Data. Complexity can walk in our world only through roads made of data. And the more data we accumulate, the more are the roads through which complexity can walk. According to a report, [6] the total amount of data in the world reached 2.8 zettabytes (ZB) in 2012 or 2.8 trillion gigabytes. The volumes of data were expected to reach 40 ZB by 2020, i.e., 4×10^{22} bytes. 90% of the world's data was generated over the last two years alone, at a rate of 2.5 quintillions (10^{18}) bytes of data a day [7, 8].

This enormous amount of data leads to a possibility, such as studying new phenomena and a problem, i.e., the need for new methodologies capable of dealing with systems with many variables. Graph theory, a field in mathematics that gained popularity in the 19th and 20th centuries, allows handling all these variables in a structure called "complex network" [9]. A graph (or network) is defined by a collection of nodes. A node can be anything: A person, an organization, a computer, a biological cell, and so forth. Two nodes may be linked, for example, because two people know each other, two organizations exchange goods, two computers have a cable connecting the two of them, or because two neurons are connected by means of synapses for passing signals. The process of modeling a phenomenon through its underlying network (whose complexity is proportional to the complexity of the original phenomenon) involves a reduction of information. It is the researcher's duty, distinguishing between necessary and unnecessary information, reducing the problem to its skeleton.

In 1999 A. L. Barabási and R. Albert published a paper titled "Emergence of Scaling in Random Networks" [10]. The authors showed that many large networks' common property is that the vertex connectivities follow a scale-free power-law distribution, that is the probability to find a node with degree k (where the degree is the number of connections of the node) in the network can be written as:

$$p(k) \propto k^{-\lambda}$$

They also argued that this is a consequence of two mechanisms: growth and preferential attachment. A new node entering the network prefers to establish a connection with a well-connected node. A model based on these two ingredients reproduces the observed stationary scale-free distributions, indicating that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

At the time of this work, we knew that many systems in nature have power-law distributions of some essential quantities. Phase-transitions, for example, have power-law distributions in the scaling parameters (the

critical exponents) that are independent of the details of the experiment. The presence of few exponents able to describe different systems justifies the epithet "universal exponents." However, phase transitions happen at only one point (in one dimension of the parameters-space). Thus the system needs a way to turn itself to the transition, and SOC (self-organized criticality) [11] proposes models where this arguably happens and extends it to an explanation of reality.

The similarity between physical systems at the criticality and systems like the Internet was a sort of insight towards a universal theory that is able to explain everything. This is why the search of ubiquitous emergent properties occurring in several different systems and transcending the specific system details is a recurrent theme in both statistical physics and complexity science [12]. With the opportunity of new discoveries, many physicists approached to phenomena different from the ones they were used to.

However, the success of this seminar work is associated to another import aspect. Together with Watts and Strogatz's work on small-world networks [13], it helped connecting scientists of different backgrounds into a new interdisciplinary field [14]. As a matter of fact, the perspective of a new, non-contaminated field led to general excitement, and avalanches of researchers approached this new discipline. Thanks to that general excitement, today we can speak about Network Science.

Since then, the applications of networks theory to real-world phenomena have been growing drastically. Today, networks play a vital role in the development of predictive models of physical, biological, and social collective phenomena [15, 16, 17]. Network's complex features often find their signature in statistical distributions, which are generally heavy-tailed, skewed, and varying over several orders of magnitude [18]. According to the finding of [10], a quite remarkable feature of many real networks is that they are approximately scale-free for sufficiently large value of k [19]. The value of the exponent λ, as well as deviations from power-law scaling, provides invaluable information on the mechanisms underlying the formation of the network, such as small degree saturation, variations in the local fitness to compete for links, and high degree

cut-offs owing to the finite size of the network [20]. Indeed real networks are not infinitely large, and the largest degree of any network cannot be larger than the number of nodes. The presence and type of such a "universal" law give insights into the driving processes or on the characteristic properties of the observed system. In particular, the power-law shape of the degree distribution, which is the hallmark of *scale-free* networks, leads to important emergent attributes such as self-similarity in the network topology, robustness to random failures, and fragility to targeted attacks. Notably, scale invariance extends far beyond the degree distribution, affecting many other quantities as weighted degree, betweenness [21] and degree-degree distance [22].

Questions concerning how a virus spreads among society are crucial today due to the current COVID-19 pandemic. The description of a spreading problem relies on the definition and characterization of a large number of individuals and their interactions in spatially extended systems. The degree of heterogeneity characterizing the topology of human interaction networks, as well as the mobility network, is one of the main culprits of the ongoing situation [23, 24]. Despite social heterogeneity and the existence of "super-spreaders" have long been known in the epidemics literature [25], a network approach added a missing piece to the existing literature. The topology of the network enters the very definition of the epidemic threshold [26, 23, 24]. In particular, larger heterogeneity levels lead to smaller epidemic thresholds (an epidemic threshold that goes to zero in the limit of infinite network). In other words, few well-connected people carrying the virus can turn a local event into a global problem.

Although this phenomenon may look different from the spreading of information in social networks, their mathematical description relies on very similar models. Appropriately modified information, "fake news" [27], allow a well-connected social-user to contaminate many other people's way of thinking. The dynamics and influence of fake news on Twitter can be so decisive to flip the results of presidential elections [28]. Notably, top spreaders can also be automated programs, also called bot [29], that amplify their influence by operating in a team [30].

If properly used, the heterogeneous structure of social networks can lead to real-time event detection. For example, Twitter data can be crucial in identifying earthquakes felt by humans and can trigger an alert typically in under two minutes. The 2014 earthquake in Napa was detected by the U.S. Geological Survey (USGS) in 29 seconds using Twitter data [31]. Therefore, it becomes of primary importance, understating what kind of information it is in the system in order to reduce false positives. While someone wants to know about an earthquake in her/his neighborhood as soon as possible, they may be less interested in the last cosmetic news sponsored by Kim Kardashian.

In these cases, Machine Learning techniques [32], in particular, natural language processes allow (prior to its training on some specific datasets) to label unstructured data, such as text (and not only). Latent Dirichlet allocation (LDA, just to mention one) is one of the most powerful data mining techniques for topic modeling [33]. Information can also be filtered by sentiment, to understand whether people are positively or negatively concerned about a topic [34]. Discussing all the machine learning applications (if this can really be done) is beyond this work's aim.

In this thesis, we retrace three important keystones of network science. Each of them can be contextualized in a different time period: the past (Chapter 2); the present (Chapter 3); and the future (Chapter 4).

The past. In the second chapter, we focus on the scale-free behavior of real-world networks. The presence or absence of a scale free-structure in real-world networks is a matter of recurrent debate in the scientific community [35]. Real-world networks are not infinitely large, and their finite dimension can hide the underlying scale-free structure in the degree distribution. Finite-size scaling [36, 37, 38, 39, 40, 41, 42], firstly developed in the field of critical phenomena and renormalization group, is a valuable tool for analyzing deviations from pure power-law behavior due to finite-size effects. We show that despite the essential differences between networks and critical phenomena, finite-size scaling provides a powerful framework for analyzing the scale-free nature of empirical networks [20].

The present. In the third chapter, we apply state-of-the-art network

analysis to an undergoing problem: the spread of COVID-19 caused by the recently discovered SARS-CoV-2 virus. We implemented a comprehensive contact tracing network analysis to find the optimal quarantine protocol to dismantle the chain of transmission of coronavirus with minimal disruptions to society [43, 18, 44, 45]. We track billions of anonymized GPS human mobility data points from a compilation of hundreds of mobile apps deployed in Latin America [46] to monitor the evolution of the contact network of disease transmission before and after the confinements [18, 44, 16]. As a consequence of the lockdowns, people's mobility across the region decreases by $\sim$53%, which results in a drastic disintegration of the transmission network by $\sim$90%. However, this disintegration did not halt the spreading of the disease. Our analysis indicates that superspreading k-core structures persist in the transmission network to prolong the pandemic [47, 48, 49, 50]. Once the k-cores are identified, the optimal strategy to break the chain of transmission is to quarantine a minimal number of "weak links" [51] with high betweenness centrality [52, 53] connecting the large k-cores.

The future. In the fourth chapter, we combine network-analyses techniques with machine-learning to define an algorithm that is able to discover trends in society, for instance, what people think about economics, education, climate change, or in this case, political elections [54]. The starting point is the failure of traditional polls in predicting presidential election outcomes across the world. To understand the reasons behind these failures, we analyze the raw data of a trusted pollster that failed to predict, along with the rest of the pollsters, the surprising 2019 presidential election in Argentina, which has led to a major market collapse in the country. Analysis of the raw and re-weighted data from longitudinal surveys performed before and after the elections reveals clear biases (beyond well-known low-response rates) related to misrepresentation of the population and, most importantly, to social-desirability biases, i.e., the tendency of respondents to hide their intention to vote for controversial candidates. We then propose a longitudinal opinion tracking method based on big-data analytics from social media, machine learning, and network theory that overcomes the limits of traditional polls. The model

achieves accurate results in 2019 Argentinian elections.

We conclude with a discussion on the results' importance, future developments and general thoughts about the future of network science.

Chapter 2

Self-similarity and scaling of complex networks

This Chapter is based on the work " True scale-free networks hidden by finite size effects" by Serafino *et al.* [20].

2.1 Introduction

In the last decade the existence of such power laws in complex networks (but also in other areas [55], e.g., power law in language [56]) has been questioned [57]. A reason of the shift in such conclusion is in the availability of larger (and new) datasets, and especially in improved statistical methods. Recently, Broido and Clauset [35] fitted a power law model to the degree distribution of a variety of empirical networks and suggested that scale-free networks are rare. Voitalov *et al.* [58] rebutted that scale-free networks are not as rare if deviations from pure power law behavior are permitted in the small degree regime. The different conclusions may depend on very fine but critical assumptions at the basis of the statistical test for the power law hypothesis. Moreover, a crucial point that is typically ignored but represents the condition for the proper use of maximum likelihood methods is the independence of the empirical observations [59]. In this work we tackle the problem of detecting power laws

in networks from a different perspective, based on the the machinery of finite size scaling.

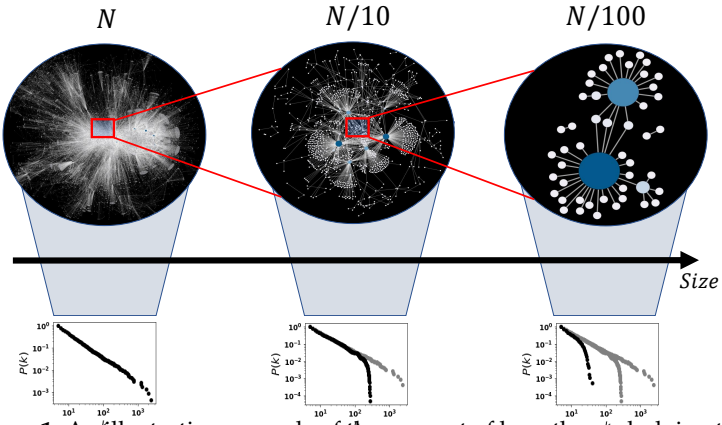


Figure 1: An illustrative example of the concept of how the underlying true scale invariance in a network may be clouded by a scale imposed by the sample size. If the degree distribution $P(k)$ of the network is scale free, then small subsamples of the network will have the same distribution (i.e., the degree structure of the network will not be altered apart from deviations at high values of k where the cutoff because of sample size operates). Specifically, the $P(k)$ vs. k log–log plots show the largest sample (Left); a reduced sample (Center), where for comparison the largest distribution is shown via gray dots; and the smallest subsample, where the two previous distributions are shown for comparative purposes (gray dots; Right). Any Anderson–Darling-like test of the sample being drawn from a scale-free distribution would fail. The network in this example is a snapshot of the structure of the internet at the level of autonomous systems [60].

Statistical physics of critical phenomena teaches us that a system at criticality exhibits power law singularities of physical quantities such as, for example, the compressibility, the specific heat, the density difference between the liquid and vapor, as well as the latent heat. Water at its critical point exhibits fluctuations at all scales between the molecular length scale and the size of the container, which could be macroscopically large. Moreover, one finds thoroughly mixed droplets of water and bubbles of gas. Indeed, any large part of the system looks like the whole – the system is self-similar. The length scale of these droplets and bubbles extends

from the molecular scale up to the correlation length, which is a measure of the size of the largest droplet or bubble. The divergence of the correlation length in the vicinity of a phase transition at the thermodynamic limit thus suggests that properties near the critical point can be accurately described within an effective theory involving only long-range collective fluctuations of the system. However, both in experiments and in numerical simulations, the infinite size limit cannot be reached and thus one observe deviations from the predicted thermodynamics limit behavior. The finite size scaling (FSS) ansatz has been developed precisely to infer the singular behavior (i.e., the exponents determining the universality classes) of the physical properties of a system in the thermodynamic limit, having only information on the system properties at finite sizes.

FSS has yet a more general validity and does not require the existence of a phase transition or an evolution process. Indeed, even though it was initially used to study finite systems near the critical point of the corresponding infinite system, FSS can be actually applied to describe structures that are self-similar when observed in a certain range of scales. As an example, we consider a Cantor set where we stop the procedure to divide intervals in three parts and removing the middle one at a scale $s_0 = 3^{-m}$. This corresponds to a fractal structure on scales between s_0 and 1, and to a non-fractal structure on scale smaller than s_0. If we measure the total length, $L(s)$, of the set with a stick of length $s = 3^{-n}$ we find $L(s) = s^{1-D}F(s/s_0)$ where $F(x) = 1$ when $x > 1$ whereas $F(x) = x^{1-D}$ when $x < 1$ and $D = \log_3 2$ is the Hausdorff–Besicovitch (or fractal) dimension of the Cantor set. Another illustration of FSS analysis is given by the truncated geometrical series $S(x, N) = \sum_0^{N-1} x^n$. When x is close to 1 it is easy to see that $S(x, N) = t^{-1}F(tN)$, where $t = 1 - x$ and $F(z) = 1-e^{-z}$. As a matter of fact, the FSS approach has been used to test scale invariance (and self similarity) also for non-critical systems such as (just to mention some very famous examples) polymers in confined geometries [61] and interfaces [62, 63]. In view of the above, FSS can also be implemented on well-established models of scale-free networks (like e.g. the Barabási-Albert model [10] or the Bak-Tang-Wiesenfeld toy model

of self-organized criticality [64]) where the scale-free behavior is not an emergent property at a critical point. Whether or not the same hypotheses holds for real world network does not undermine the possibility of applying FSS to them.

Employing the FSS machinery to test whether empirical networks display scale-free behavior in their degree distribution is not straightforward though. Unlike for physical systems, representations of a network at different scales are typically not available. Thus, in order to test whether a network shows a power law distribution of its degree k, we have constructed smaller size subsamples, effective representations of the underlying population, drawn in an unbiased manner. We then use the characteristics of the large original network as well as the derived sub-networks to test the scale-free hypothesis. Figure 1 shows an illustration of this procedure for a snapshot of the structure of the Internet at the level of autonomous systems [60]. Section 2.2 provides a brief summary of finite size scaling applied to network topology. Section 2.3 present an independent method of determining whether networks are scale-free based on analyses of the size dependence of the ratio of moments of the degree distributions. Section 2.4 provides information on the sampling scheme used to build sub-networks and on the region selected for the scaling analysis.

In the Results section we test the scale-free hypothesis, (the power law behavior in the degree distributions) on around two hundred large empirical networks (those considered in [35] and [58]). Remarkably, we find that such a venerable hypothesis cannot be rejected for many (but not all) networks. Moreover the two scaling exponents for such networks satisfy an additional scaling relationship, which derives from the shape of the degree cross-over in scale-free networks. We benchmark our results against the quality measure of the well-known scale-free graph introduced by Barabási and Albert [10]. Further we show that finite size scaling allows discerning pure power laws from log-normal and Weibull distributions. In conclusion, our results support the claim that scale invariance is indeed a feature of many real networks, with finite size effects accounting for quantifiable deviations. We conclude the chapter with a

discussion of the results.

2.2 Finite Size Scaling of networks

A scale-free network is postulated to have a degree distribution $p(k) \propto k^{-\lambda}$ beyond some lower degree cut-off k_{min}. For an infinitely sized network, since $k_{min} \geq 1$, the exponent $\lambda > 1$ in order for $p(k)$ to be normalizable. In what follows, we will consider the cumulative distribution $P(k) = \int_k^\infty p(q)\, dq \propto k^{-\gamma}$ where $\gamma = \lambda - 1 > 0$.

Networks are of course not infinitely large. In a network comprising N nodes, k can be at most equal to $N - 1$. This is the intrinsic limit on k given by the network size. Thus it is plausible that, below some k_c (cross-over value), the degree distribution follows a power law behavior as would be expected for an infinite network but falls more rapidly beyond k_c. The finite size scaling hypothesis states that

$$P(k, N) = k^{-\gamma} f(kN^d) \tag{2.1}$$

where $d < 0$. The remarkable simplifying feature of the scaling hypothesis is that P is not an arbitrary function of the two variables k and N but rather k and N combine in a non-trivial manner to create a composite variable. The behavior of the system is fully defined by the two exponents, γ and d, and the scaling function f. The exponent $d < 0$ so that, for an infinite size network ($N \to \infty$), the argument of f approaches zero. A pure power law decay of $P(k, N)$ with k for very large N requires that $f(x) \to$ constant as $x \to 0$. The additional normalization condition is $f(x) \to 0$ sufficiently fast when $x \to 1$. The finite size effects are quantified by the behavior of the function f as its argument increases, e.g., when $k \gtrsim k_c$. For a network with a finite number of nodes, the degree distribution does not follow a pure power law but is modified by the function f (see also [65] for a discussion of finiteness in the context of growing network models).

A powerful way of assessing whether a network is scale invariant is to confirm the validity of the scaling hypothesis and determine the

two exponents and the scaling function f by using the collapse plot technique. One may recast Eq. (2.1) as

$$P(k,N)k^{\gamma} = f(kN^{d}). \tag{2.2}$$

Then the path forward is simple. For networks belonging to the same class but with different N, one optimally selects two fitting parameters γ and d by seeking to collapse plots of $P(k,N)k^{\gamma}$ versus kN^{d} for different N on top of each other [66]. The fidelity of the collapse plot provides a measure of self-similarity and scale-free behavior, the optimal parameters are the desired exponents, and the collapsed curve is a plot of the scaling function.

We start out with a single representation of an empirical network with N nodes. For purposes of the scaling collapse plot, we seek additional representative networks of smaller sizes. In order to accomplish this, we obtained the mean degree distributions of multiple sub-networks of sizes $\frac{N}{4}$, $\frac{N}{2}$ and $\frac{3N}{4}$, which were then collapsed on to each other and the original network to create a master curve. The quality S of the collapse plot is then measured as the mean square distance of the data from the master curve in units of standard errors. S is thus like a reduced χ^2 test, and should be around one if the data really collapse to a single curve and much larger otherwise [67].

Note that as a measure of the size of a network (or sub-network), one may use the number of nodes N or alternatively the number of links E. The scaling function in this case reads as follows:

$$P(k,E)k^{\gamma} = f_{E}(kE^{d_{E}}), \tag{2.3}$$

where the exponent γ is the same as before and the exponent $d_E < 0$ ought to be equal to the previously introduced exponent d for networks satisfying the finite size scaling hypothesis (see next section).

2.2.1 Quality of collapse

We now describe the procedure for deriving the master curve of the scaling function from the cumulative degree distributions of the various sub-networks, following the steps described in [67, 68]. The key premise is

that when these distributions are properly rescaled they can be fitted by a single (master) curve. The quality of the collapse plot is then measured as the distance of the data from the master curve, and the collapse is good if all the rescaled distributions overlap onto each other.

In practice for each (sub-)network size $n \in \{\frac{N}{4}, \frac{N}{2}, \frac{3N}{4}, N\}$ we have the set $\{j\}$ of ordered points for the cumulative degree distribution in the form $\{(k_j, P(k_j, n))\}_j$. After applying the scaling laws we have:

$$\begin{cases} x_{nj} = k_j \, n^d \\ y_{nj} = P(k_j, n) \, k_j^\gamma \end{cases}$$

so that x_{nj} is the rescaled j^{th} degree in the distribution of the n-sized sub-network, and y_{nj} is the rescaled value of such distribution relative to the j^{th} degree. We also assign an error on the latter quantity as $dy_{nj} = dP(k_j, n) \, k_j^\gamma$, where $dP(k_j, n)$ is the Poisson error on the count $P(k_j, n)$ — see Appendix A.

The master curve Y is the function best fitting all these points. We define the quality of the collapse as

$$S = \frac{1}{3|M|} \sum_{(n,j) \in M} \frac{(y_{nj} - Y_{nj})^2}{dy_{nj}^2 + dY_{nj}^2}, \tag{2.4}$$

where Y_{nj} and dY_{nj} are the estimated position and standard error of the master curve at x_{nj}, while M is the set of terms of the sum (roughly, the set of points for which the curves for the various n overlap).

For each x_{nj}, in order to define Y_{nj} and dY_{nj} we first need to select a set of points m_{nj} as follows. In each of the other sets $n' \neq n$, we select (and put in m_{nj}) the two points j' and $j' + 1$ that best approximate x_{nj} from below and above, i.e., the two points such that $x_{n'j'} \leq x_{nj} \leq x_{n'(j'+1)}$. If this procedure fails to select two points for each $n' \neq n$, then Y_{nj} and dY_{nj} are undefined at x_{nj} which thus does not contribute to S (this happens if set n is alone in this region of x and is the master curve by itself). Otherwise, we compute Y_{nj} and dY_{nj} using a linear fit through the selected points in $(n', l) \in m_{nj}$, so that Y_{nj} is the value of that straight

line at x_{nj} and dY_{nj} is the associated standard error:

$$Y_{nj} = \frac{W_{xx}W_y - W_x W_{xy}}{\eta} + x_{nj}\frac{WW_{xy} - W_x W_y}{\eta} \tag{2.5}$$

$$dY_{nj}^2 = \frac{1}{\eta}(W_{xx} - 2x_{nj}W_x + x_{nj}^2 W) \tag{2.6}$$

where $w_{n'l} = 1/dy_{n'l}^2$ for the fit weights and

$$\begin{aligned} W &= \sum_{(n'l)\in m_{n'j}} w_{n'l} \\ W_x &= \sum_{(n'l)\in m_{n'j}} w_{n'l}x_{n'l} \\ W_y &= \sum_{(n'l)\in m_{n'j}} w_{n'l}y_{n'l} \\ W_{xx} &= \sum_{(n'l)\in m_{n'j}} w_{n'l}x_{n'l}^2 \\ W_{xy} &= \sum_{(n'l)\in m_{n'j}} w_{n'l}x_{n'l}y_{n'l} \\ \eta &= WW_{xx} - W_x^2 \end{aligned} \tag{2.7}$$

for the fit parameters.

The quality of the collapse S measures the mean square distance of the sets to the master curve in units of standard errors, analogously to a χ^2 test [67]. The number of degrees of freedom can be estimated by noting that each of the $|M|$ points of the sum of S has in turn 3 intrinsic degrees of freedom: $|m|$ points as described above (6 in our case) minus 2 from computing mean and variance of Y, minus 1. Hence by using $3|M|$ as normalization factor, S should be around one if the data really collapse to a single curve and much larger otherwise.

We optimize the quality S of the collapse by varying the scaling exponents γ in the interval $\Gamma - 0.5 \leq \gamma \leq \Gamma + 0.5$ and d in the interval $d - 0.1 \leq \gamma \leq d + 0.1$. The errors associated with γ and d are estimated with a $S + 1$ analysis: $\Delta\gamma$ is such that $S(\gamma + \Delta\gamma) = S(\gamma) + 1$ and Δd is such that $S(d + \Delta d) = S(d) + 1$.

2.3 Ratio of moments test

A simple alternative and independent test of the scale-free hypothesis is to study the size dependence of the ratio between the $i-$th and the $(i-1)-$th moments of k, for various i. The $i-$th moment $\langle k^i \rangle$ is defined to be

$$\langle k^i \rangle = \int_{k_{min}}^{\infty} k^{i-1} k^{-\gamma} f(kN^d)\, dk \propto N^{-d(i-\gamma)} \tag{2.8}$$

provided $i > \gamma$. Instead if $i \leq \gamma$, $\langle k^i \rangle$ converges to a constant value for $N \to \infty$. Therefore when $i - 1 > \gamma$,

$$\langle k^i \rangle \Big/ \langle k^{i-1} \rangle \propto N^{-d}, \tag{2.9}$$

independently of i. Thus, for a scale-free network, a log-log plot of the ratio of consecutive moments versus N is a straight line with slope $-d$. Likewise

$$\langle k^i \rangle = \int_{k_{min}}^{\infty} k^{i-1} k^{-\gamma} f_E(kE^{d_E})\, dk \propto E^{-d_E(i-\gamma)} \tag{2.10}$$

when $i > \gamma$, otherwise $\langle k^i \rangle$ goes to a constant for $E \to \infty$. Therefore when $i - 1 > \gamma$,

$$\langle k^i \rangle \Big/ \langle k^{i-1} \rangle \propto E^{-d_E}. \tag{2.11}$$

The exponents d and d_E are not independent for scale-free networks. On the one hand, Eqs. (2.8) and (2.10) imply $E \propto N^{d/d_E}$. On the other, in general $\langle k \rangle \propto E/N \propto N^{d/d_E - 1}$. Due to the above equations $\langle k \rangle$ is constant for scale-free networks with $\gamma > 1$, implying that $d = d_E$. Thus the difference between d and d_E values (that we statistically assess through their $Z-$score) provides an independent quality measure of the scale-free attributes of a network.

2.4 Sub-sampling and scaling region

In order to generate a sub-network of a given size $n < N$, we pick n nodes at random among the N nodes of the original network, removing all the other nodes and the links originating from them. It is well known

that the sub-sampling procedure modifies the shape of the degree distribution of the network. In particular, sub-networks of scale-free networks are not scale-free because of deviations at low k values [69] (this happens independently of the sampling scheme adopted [70]). The problem of the left tail of the distribution however applies more generally, because deviations from the scale-free behavior at low degrees are rather common in empirical and network models. Therefore we perform the scaling analysis described above only for $k \geq k_{min}$, where the lower bound of the scaling region k_{min} is chosen such that the empirical distribution of the original network and its best power law fit (with exponent Γ, computed with the maximum-likelihood method of Clauset, Shalizi and Newman [57], see Appendix A) are as similar as possible above k_{min} [71]. In the Appendix A we show that this allows us to get rid of any deviations induced by the sub-sampling scheme. However, when the empirical distribution of the network deviates substantially from a power law over its entire domain, then the estimated k_{min} can become very large and may even diverge. In these cases the number of nodes n^* of the (sub-)network with $k \geq k_{min}$ becomes very small or vanishing, yielding an unstable or undefined collapse. We thus use $n^* \geq \ln N$ as a condition on as the minimum number of nodes in each (sub-)network for the feasibility of the scaling analysis.

2.5 Results

To sum up, two independent statistical tests of the scale-free attributes of a network explained in subsections A and B are the quality of the collapse S (i.e., the reduced χ^2 between data and master curve) and the compatibility of d and d_E (measured through their $Z-$score). Figure 2 outlines the flow of the analysis. In line with Broido & Clauset [35] and Voitalov *et al.* [58], we use these tests to define a classification for the degree distribution of empirical networks:

- **SSF** (strong scale-free) if $S \leq 1$ and $Z_{dd_E} \leq 1$,
- **WSF** (weak scale-free) if $S \leq 3$ and $Z_{dd_E} \leq 3$,

- **NSF** (non scale-free) otherwise or when $n^* < \ln N$ for the original network or any of its sub-networks.

Note the nestedness of the classification, for which a SSF network is also WSF.

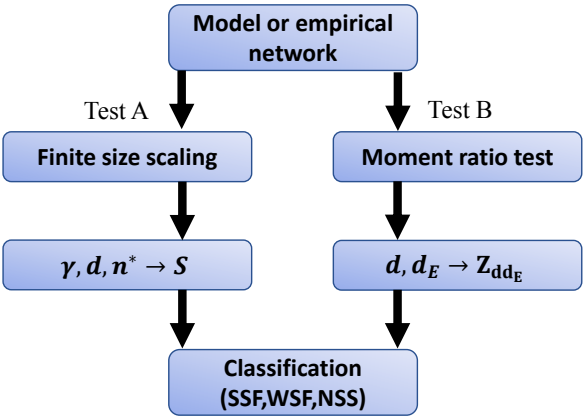


Figure 2: Schematic flow of the analysis. For a given network we independently perform two tests: the finite size scaling and the moment ratio test. The respective statistical outputs (the quality of the collapse S and the Z-score between d and d_E) are combined to obtain a classification of the networks as **SSF**,**WSF** or **NSF**.

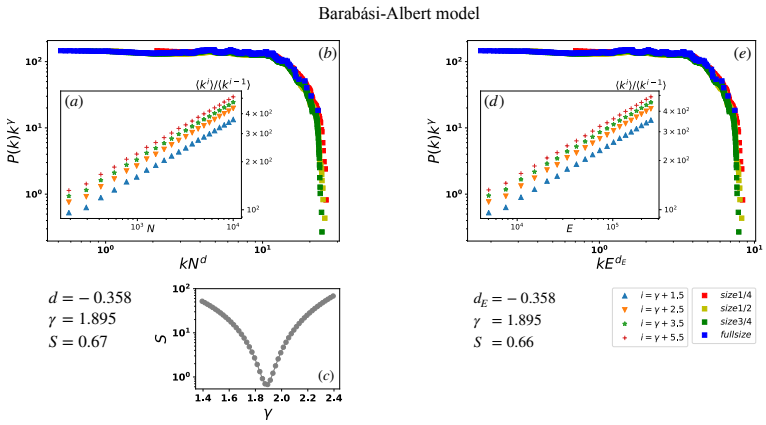


Figure 3: Scaling analysis on a numerical realization of the Barabási-Albert model. The network has $N = 10^4$ nodes and the minimum node degree is $k_{min} = 14$. The best power law fit on this network yields $\Gamma = 1.89 \pm 0.02$. Note this value is smaller than $\Gamma = 2$ because of deviations from the pure power law at small ks: indeed, the theoretical $P(k)$ in the Barabási-Albert model goes as $[k(k+1)(k+2)]^{-1}$ [72]). Panels (a), (b), (c) show results of the scaling analysis using the number of nodes as for Eqs. (2.2) and (2.9). Inset (a) reports the dependence of various moment ratios on N; fitting these slopes yields $d = -0.358 \pm 0.035$. The main panel (a) shows the collapse of the cumulative degree distributions when scaled with N. The best collapse is obtained with $\gamma = 1.89 \pm 0.06$ and yields $S = 0.67$. Panel (c) shows how the quality of the collapse reported in (a) varies on moving away from the optimal value of γ. Panels (d), (e) further show results of the scaling analysis using the number of links as for Eqs. (2.3) and (2.11). In this case, the moment ratio test of inset (d) returns $d_E = -0.351 \pm 0.031$ while the best collapse of the cumulative degree distributions reported in the main panel (e) is obtained with $\gamma = 1.89 \pm 0.05$ and yields $S = 0.66$.

2.5.1 Power law and Poisson distribution

We start analyzing the reference cases of Barabási-Albert [10] and Erdős-Rényi [73] models whose behavior is known. In the former case $p(k) \sim k^{-3}$, whereas, in the latter case $p(k) \sim \text{Poisson}_{\bar{k}}(k)$. Figure 3 shows that for a realization of the Barabási-Albert graph the degree distributions of

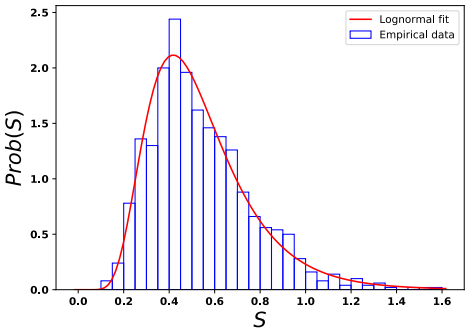


Figure 4: Empirical distribution of the quality of collapse S obtained from finite size scaling analysis on 1000 realizations of the Barabási-Albert graph (same parameters of Figure 3). The distribution is well fitted by a log-normal with $\mu = -0.70 \pm 0.1$ and $\sigma = 0.414 \pm 0.009$.

the (sub-)networks result in a collapse of very high quality. The power law exponent γ yielding the best collapse is consistent with the value Γ obtained by maximum-likelihood fitting the degree distribution of the mother network with a power law [57]. Additionally, the moments ratio are indeed parallel lines, with compatible slopes d and d_E. A more robust statistics is obtained by analysing 1000 realizations of the Barabási-Albert model (Figure 4). Within this sample, 98% of the networks are classified as SSF while 2% as WSF. The estimated scaling exponents are all consistent with each others among the different realizations.

For the Erdős-Rényi model the estimated k_{min} for the degree distribution is so large that it is not possible to have (sub-)networks with number of nodes $n^* \geq \ln N$ (in principle, for this network, the k_{min} estimated from the KS test should be larger than the largest degree of the network). As such, the Erdős-Rényi graph is classified as NSF. We obtained the same outcome in an ensemble of 1000 realization of this network model.

2.5.2 Alternative fat tail distributions

While the power law is the only distribution featuring scale invariance, there are other distributions characterized by a fat right tail that can re-

semble a power law in finite systems. Hence determining which of these distribution better fits empirical network data is often a nontrivial task. In particular the classical approach based on p-values computed from a Kolmogorov-Smirnov test (see Appendix A) is able to rule out some competing hypothesis but not to confirm one [57]. Moreover, the hypothesis testing approach may fail when applied to regularly varying distributions [58]. It is therefore meaningful to put our finite size scaling approach to the test of alternative fat tail distributions. Here we consider the representative cases of the log-normal and Weibull distributions. The log-normal distribution $p(\ln k) = \text{Normal}(\mu, \sigma)$ is characterized by parameters μ and σ, respectively the mean and standard deviation of the variable's natural logarithm. For large values of σ this distribution is highly skewed and features a fat tail for large k values. The Weibull distribution $p(k) = (h/l^h)k^{h-1}\exp\left[-(k/l))^h\right]$ is characterized by parameters h (shape) and l (scale). The fat tail in this case appears for $h \to 0$. We use the Viger-Latapy algorithm [74] to generate networks with these degree distributions.

Figure 5 shows the scaling analysis for a realization of a network with log-normal $p(k)$ and for another realization with Weibull $p(k)$. In both cases we observe that the quality of the collapse is poor and that the moment ratios are not parallel lines. Therefore both networks are classified as NSF. Moreover, S as a function of γ does not show any minimum in the region around Γ (the minimum does exist, but is located elsewhere). This means that the exponent estimated by finite size scaling γ and that obtained from maximum likelihood power law fitting Γ are substantially different: the outcome of the scaling analysis is not consistent in this case. However, the result depends much on the choices of parameters characterizing the distribution. Indeed Figure 6 shows that the percentage of networks classified NSF decreases by increasing σ in the log-normal case, as well as by decreasing h in the Weibull case – up to a point where the variance of the distributions becomes so large that the scaling analysis can hardly distinguish these distributions from power laws at finite N. For these cases, the value of γ that minimizes S is indeed compatible with Γ.

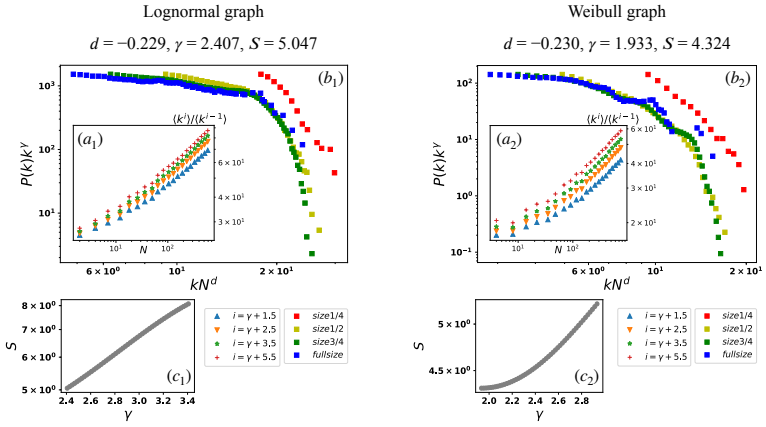


Figure 5: Scaling analysis (with N) on a numerical realization of a log-normal graph with $(\sigma, \mu) = (0.8, 1.8)$ (panels a_1, b_1, c_1) and of a Weibull graph with $(h, l) = (0.6, 1.8)$ (panels a_2, b_2, c_2). In both cases the network has $N = 10^4$ nodes. Log-normal graph: the best power law fit is obtained with $\Gamma = 2.90 \pm 0.12$, the moment ratio tests yield $d = -0.209 \pm 0.033$ and $d_E = -0.208 \pm 0.033$, and the best collapse is obtained with $\gamma = 2.40 \pm 0.37$ and yields $S = 5.047$. Weibull graph: the best power law fit is obtained with $\Gamma = 3.43 \pm 0.08$, the moment ratio tests yield $d = -0.230 \pm 0.037$ and $d_E = -0.219 \pm 0.036$, and the best collapse is obtained with $\gamma = 1.933 \pm 1.055$ and yields $S = 3.271$.

2.5.3 Real world networks

At last we move to real network data. We consider a large set of empirical networks taken from the Index of Complex Networks (ICON) as well as from the Koblenz Network Collection (KONECT). These are the datasets used by Broido & Clauset [35] and Voitalov *et al.* [58]. See the Appendix A for a discussion on how we built the dataset. Overall, we have networks belonging to ten different categories: biological (PPI), social (i.e., friendship and communication), affiliation, authorship (including co-authorship), citation, text (i.e., lexical), annotation (i.e., feature, folksonomy, rating), hyperlink, computer, infrastructure. Figure 7 shows results of the finite size scaling analysis for selected network instances,

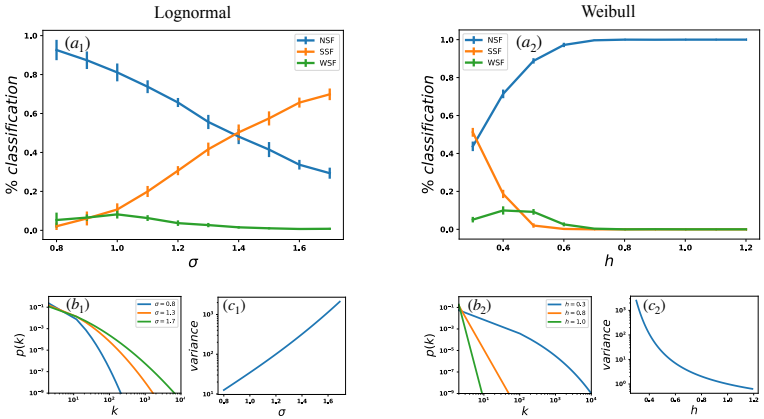


Figure 6: Outcome of the scaling analysis (with N) on log-normal and Weibull networks as a function of the parameters of the degree distributions, respectively (μ, σ) and (l, h). Panels (a_1) and (a_2) show the percentage of networks classified as strong, weak and non scale-free for varying σ at fixed $\mu = 1$, and for varying h at fixed $l = 3.5$, respectively. This statistics is computed over ensembles of 2000 networks for each choice of parameters σ and h. Panels (b_1) and (b_2) show representative instances of the distribution in the range of parameters analyzed, whereas, panels (c_1) and (c_2) displays the corresponding value of the variance of the distribution. Note that we do not report results for varying μ at fixed σ nor for varying l at fixed h, because we observe almost no dependency of the classification on these parameters.

whereas, Figure 8 and Table 1 summarize results of the scaling analysis for all the networks considered. The main outcomes of the analysis are the following.

- Figure 8(a): the scaling exponents d and d_E obtained from the moment ratio test are compatible in most of the cases.
- Figure 8(b): the value of γ computed from finite size scaling is often in good agreement with Γ obtained from the maximum likelihood power law fit of the degree distribution [57].
- Figure 8(c): the exponents γ and d of the scaling function are not

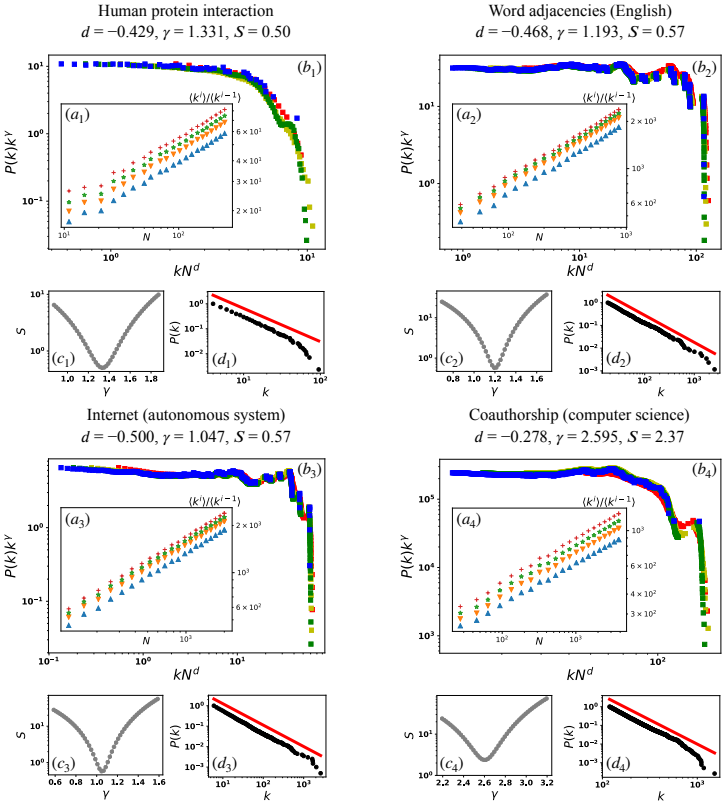


Figure 7: Scaling analysis (with N) on four real network instances. Top left panels (1): the 2005 version of the proteome-scale map for Human binary protein-protein interactions ($N = 1706$, $E = 3155$) [75]. Top right panels (2): the word adjacency graph extracted from the English text "The Origin of Species" by C. Darwin ($N = 7724$, $E = 46281$) [76]. Bottom left panels (3): (symmetrized) snapshot of the Internet structure at the level of Autonomous Systems in 2007 ($N = 26475$, $E = 53381$) [77]. Bottom right panels (4): the collaboration graph of authors of scientific papers from DBLP computer science bibliography ($N = 1314050$, $E = 10724828$) [78]. Panels (a), (b), (c) are analogous to those reported in Figures 3 and 5, whereas, panels (d) visually show the classical plots of $p(k)$ in double logarithmic scale together with the plot of the estimated slope γ using FSS analysis.

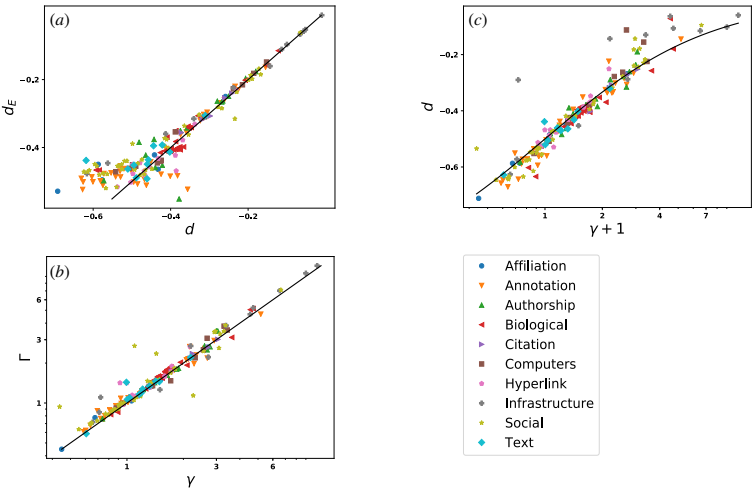


Figure 8: Visual summary of results from the finite size scaling analysis, in which each network dataset is represented as a point in a specific plane. Panel (a) shows the relation between d and d_E resulting from the moment ratio test, with the solid black line representing the identity. The other two panels refer to the scaling analysis with N. Panel (b) shows the relation between γ computed from finite size scaling and Γ from the maximum likelihood power law fit of the degree distribution (see Appendix A). The solid line again represents the identity. Panel (c) shows the relation between the exponents γ and d of the scaling function, with the solid black line representing the curve $d = -(\gamma + 1)^{-1}$ (the text has details).

independent but satisfy a universal relation $d \simeq -(\gamma + 1)^{-1}$, which derives from the nature of the degree cross-over in scale-free networks – namely the maximum degree for which the power law behavior holds. According to Eq. (2.1), this is the value k_c for which the scaling function $f(x) \to 0$ (graphically speaking, when the master curve $P(k)k^\gamma$ falls down), corresponding to $x \gtrsim 1$ whence $k_c \sim N^{-d}$. The analysis presented in Figure 8(c) suggests that $k_c \sim N^{1/(\gamma+1)}$, and in agreement with theoretical results we find that also the maximum degree of the network k_{max} scales in the same way (see Appendix A). However this scaling behavior is somehow

different from the $k_c \sim N^{1/\gamma}$ as predicted by hand-waving argument [79, 80, 81], likely due to inner correlations in the networks which modify the value of the cross-over [80].

- No particular relation between quality of collapse S and estimated exponent γ is found, nor any clusterization of networks amenable to categories within the plane defined by these two variables (see Appendix A). However this result is obtained when the different network categories are well balanced in the dataset, because networks that are very similar tend instead to cluster together. This is for instance the case of protein interaction networks belonging to different species. In order to remove this artificial clustering effect, we have not considered in our dataset these (and other) cases of very similar networks nor repetitions of the same network (see Appendix A). This is the main reason why our dataset is apparently smaller than that used by Broido & Clauset [35].

- Overall, as shown in Table 1, the 185 networks of our dataset are classified as strong scale-free (SSF) in the 27% of cases, weak scale-free (WSF) for the 23% and non-scale-free (NSF) for 50%. This classification however does vary substantially among the different network categories. On the one hand, biological networks are very often classified at least as WSF. The same happens for computer and hyperlink networks, with outliers respectively given by the Gnutella peer-to-peer file sharing network (that has the same character of a social networks [82]) and by some hyperlink networks restricted to specific domains. Citation and text networks are few in our analysis, but are often scale-free. On the other hand, infrastructure networks (i.e., road and flights network) are rarely scale-free (with the notable exception of Air traffic control systems), possibly because of the heavy cost of establishing a connection. Between these two extremes, there are the social and other kinds of networks (see for instance the well-known discussion of the Facebook case presented in [83, 84], and that of other information sharing social network presented in [85]).

Type	Number	SSF(%)	WSF(%)	NSF(%)
Affiliation	8	63	12	25
Annotation	38	21	24	55
Authorship	15	27	20	53
Biological	30	40	30	30
Citation	5	40	0	60
Computer	13	39	38	23
Hyperlink	14	22	21	57
Social	39	13	18	69
Text	11	55	27	18
Infrastructure	12	0	17	83
Total	185	27	23	50

Table 1: Classification of empirical networks (split into categories). For each category we report the total number of networks and the percentage of SSF, WSF and NSF instances. For detailed results on each network analyzed, see the Supplementary Dataset Table.

2.6 Discussion

Since the onset of network science, scale invariance of complex networks has been regarded as a universal feature present in real data [86, 87, 88, 89, 90, 57] as well as reproduced in models [10, 91, 92, 72, 93, 94]. Thus the recent claim by Broido & Clauset [35] that scale-free networks are rare created a stir, strengthening previous claims along the same direction [95, 57, 55]. Voitalov *et al.* [58] replied to these arguments fitting data to generalized power laws, that is, regularly varying distributions $p(k) = l(k)k^{-\lambda}$ (where $l(k)$ is a function that varies slowly at infinity and thus does not affect the power law tail). By allowing deviations from the pure power law distribution at low k, they argued that scale-free networks are definitely not rare. Gerlach & Altmann [59] very recently touched on this issue, showing that correlations present in the data can lead to false rejections of statistical laws when using standard maximum-likelihood recipes (in the case of networks, this can be important in the presence of degree-degree correlations).

In this work we go beyond statistical arguments and apply power-

ful tools from the study of critical phenomena in physics to analyze a wide range of model and empirical networks. Here we have showed that many of these networks spontaneously, without fine-tuning, satisfy the finite size scaling hypothesis, which, in turn, supports the claim that complex networks are inherently scale-free. While a direct comparison with the results previously discussed would be interesting, the final results would not be meaningful, given the differences in the underlying hypotheses of the different models. We have shown how different hypotheses can lead to distinct results. The hypothesis underlying our approach, which came from results previously obtained in the field of statistical mechanics and critical phenomena, goes beyond the applications they were initially designed for and does not require the existence of a critical point. Together with previous work, our methodology fits in the bag of tools that a researcher can use in order to assess the scale free character of a network.

Our scaling analysis is based on the extraction of small representations of the networks using a random node selection scheme. Of course, an intrinsic limitation of any rescaling method applied to network data is the impossibility to consider system sizes spanning orders of magnitude. As a further general remark, finding a robust method to rescale (or coarse grain [96, 97]) a network is still an open issue in the literature since networks are not embedded in any Euclidean space. Commonly used approaches lack generality since they are based on the choice of the embedding geometric space [98] or on the average path length [99]. In order to avoid *ad hoc* assumptions, we decided to follow the simplest (although not necessarily the most accurate) scheme. As shown in the Appendix A, by averaging over many extraction of the sub-network we are able to preserve the degree distribution of the original network, that is what we are interested in. Finally note our claims regards the self-similarity of the degree distribution, but we restrain ourselves in making general conclusions about the overall self-similarity of networks – this would involve the study of other quantities such as clustering, average path length and so on [100].

Chapter 3

Superspreading k-cores

This Chapter is based on the work [45], currently under submission.

3.1 Introduction

In the absence of vaccine or treatment for COVID-19, state-sponsored lockdowns have been implemented worldwide to halt the spread of the ongoing pandemic creating large social and economic disruptions [101, 102, 103]. In addition, some countries have also implemented digital contact tracing protocols to track the contacts of infected people and reinforce quarantines by targeting those at high risk of becoming infected [104, 105, 106, 107, 108, 109, 110, 111, 112, 113]. Here we develop, calibrate, and deploy a contact tracing algorithm to track the chain of disease transmission across society. We then search for intelligent quarantine protocols to halt the epidemic spreading with minimal social disruptions [18, 43, 44, 114, 115, 116].

Our study uses two complementary datasets. The first includes data from "Grandata-United Nations Development Programme partnership to combat COVID-19 with data" [46]. It is composed of anonymized global positioning system (GPS) data from a compilation of hundreds of mobile applications (apps) across Latin America that allow to track the trajectories of people (users). The data identify each mobile phone

device with a unique encrypted mobile ID and specifies its latitude and longitude location through time, encoded by geohash with 12 digits precision. Typically, this dataset generates $\sim$ 450 million data points of GPS location per day across Latin America in particular in the state of Ceará, Brazil (see Appendix B.1).

The second dataset is an anonymized list of confirmed COVID-19 patients obtained from the Health Department authorities from both states. It includes the geohash of the address, the SARS-COV-19 test detection date and first day of symptoms of COVID-19 (see Appendix B.1). This information is crucial to determine the time period when the patient was contagious.

We cross-match the geolocation of the patients with the GPS dataset obtaining the encrypted mobile ID of the patients. We then trace the geolocalized trajectories of COVID-19 patients during a period -14/+7 days from the onset of symptoms to look for contacts of the infected person to define the transmission network using the model described below (see Appendix B.1).

When it comes to GPS data, the definition of contact is not clear. Two persons with the same GPS coordinate may never have met if the time constraints are not satisfied, i.e., if they were in such a position at different hours of the day. In section 3.2 we introduce a probabilistic definition of a contact. Let us notice that in this work we only focus between infected/potentially infected persons and healthy persons. With this definition of contact in mind we build a contact network and we monitor its structural properties over time (section 3.3). We present the main results in section 3.4. We conclude the chapter with a discussion of the results (3.5). Moreover, we propose a practical application of our algorithm.

3.2 Contact model

The COVID-19 spreading model is represented by a Susceptible-Exposed-Infectious-Recovered (SEIR) process [18], Figure 9a. Infectiousness starts 2 days before and lasts up to 5 days after the onset of symptoms [117]. In this work, we add two days to each of these limits to conservatively

capture most transmissions. Thus, in principle, to trace those people potentially infected by COVID-19 patients, we track contacts 4 days before and 7 days after the reported date of first symptoms, Figure 9a. In addition, we extend the tracing period back in time to also consider exposures that could come from asymptomatic cases. Exposures start the incubation period of the infected person which can occur up to 12.5 days before onset of symptoms (5.2 days on average, 95% percentile 12.5 days [118, 119], Figure 9a). To conservatively trace exposure events, we add $\sim$2 days to this incubation period and obtain the widely used 14 days period. Hence, to trace transmission and exposure cases, we perform contact tracing over -14/+7 days from onset of symptoms, (Figure 9a. We note that the peak of infectiousness as well as 44% (95% confidence interval, 25-69%) of infected cases occur during the pre-symptomatic stage [117]. Thus, performing contact tracing is essential to stop the spreading disease.

3.2.1 A probabilistic interpretation

The GPS geolocation of the trajectories of both infected and susceptible people is used to trace several layers of contacts in the transmission network using the following model. A contact at time stamp n is initiated with an infected user (source) at time t_0, as it is shown in Figure 9b. At t_0 we draw a contact area as a circle centered in the source position with a radius r. We then gather all the GPS datapoints from susceptible users (targets) that enter the contact area from t_0 to $t_0 + T$, where T is the total exposure time. We follow the trajectories of source and target within the time-space area and compute the probability of infection at time stamp n as $p_i[n] = p_d[n] \cdot p_t[n]$. The first component of the probability of infection is the space component [44]:

$$p_d[n] = 1 - \frac{d[n]}{2r} \tag{3.1}$$

with $d[n] = |\langle s \rangle - \langle t \rangle|$, where $\langle x \rangle$ in $\{s, t\}$ refers to the average position of source/target inside the time-space area. That is, it is the distance

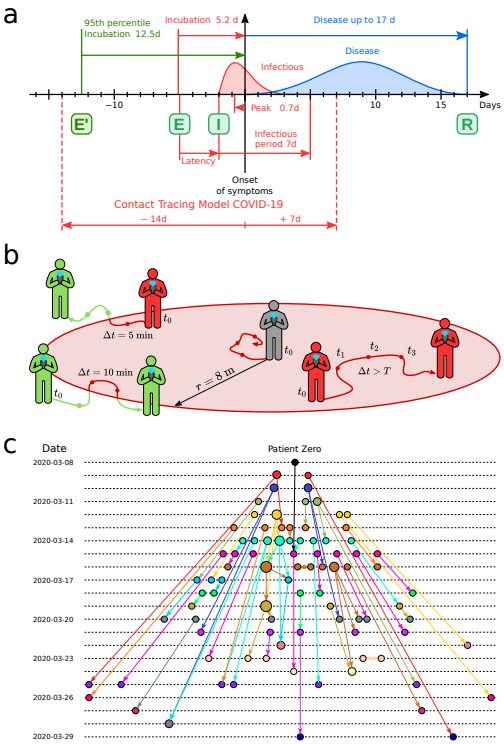


Figure 9: (a) Infectiousness profile of COVID-19. The COVID-19 pandemic is represented with a SEIR model. From exposure (E) the virus is incubated in average for 5.2 days (12.5 days 95^{th} percentile), starting the symptoms 2 days after infectiousness (I) and lasting the disease up to 17 days to recover (R). We use a window -14/+7 days from the first symptoms to detect infectious and exposure. **(b)** Contact area used in the contact tracing model. The grey person is at the first datapoint of the source at t_0. We collect all datapoints for every user in a T=30 min forward window $(t_1, t_2, t_3, ..., t_0 + T)$ within an 8 m circle from the initial position. For each target (green and red) we compute the average position and the time spent inside the contact area (red part of the trajectory line). **(c)** Partial transmission tree of outbreak of confirmed SARS-CoV-2 infection identified by contact tracing during calibration in the month of March 2020. Links goes from the source of infection to the target. The colors represent the day of first symptoms for each node and size is the out-degree.

between the source and target average positions of the data points within the contact area.

The second component (the time component) is proportional to the overlapped amount of time that source and target spent within the contact area:

$$p_t[n] = \frac{\tau(\Delta t^s, \Delta t^t)[n]}{T}, \tag{3.2}$$

and

$$\tau(\Delta t^s, \Delta t^t)[n] = \begin{cases} 0, & \text{if } \max(t_f^t, t_f^s) > \min(t_l^t, t_l^s) \\ \max(t_f^t, t_f^s) - \min(t_l^t, t_l^s), & \text{otherwise.} \end{cases} \tag{3.3}$$

Each user needs at least two data points within the contact area to define its Δt^x with x in $\{s, t\}$, i.e. the amount of time source/target spent in the time-space area. If this condition is not met, $\tau(\Delta t^s, \Delta t_t)[n] = 0$.

According to the previous definitions, when the average distance between source and target is zero, then $p_d[n] = 1$, and when the average distance is $2r$, then $p_d[n] = 0$. On the other hand, when the exposure time $\geq T$, then $p_t[n] = 1$, and decreases to $p_t[n] = 0$ as the exposure time decreases. The probability $p_d[n]$ quantifies the contact probability for two users in the same area defined by r. A contact requires non only a space overlapping but also a time overlap, $p_t[n]$, which quantifies the probability that two users met based on the time commonly spent in the same area. We then combine these two probabilities for each timestamp n into their product:

$$p_i[n] = p_d[n] \cdot p_t[n] \tag{3.4}$$

Contacts with low probability of infection $p_i[n]$, but repeated throughout time, can also infect the target. To incorporate this effect in the model, we define the probability of infection for a series of repeated contacts $P_i[n]$ as a recursive formula from time 1 to n with $P_i[0] = 0$:

$$P_i[n] = p_i[n](1 - P_i[n-1]) + P_i[n-1]. \tag{3.5}$$

The iteration of contacts between source and target, $P_i[n]$, generates higher probability of infection than a single contact $p_i[n]$. This means that there is a difference between a short single contact between two people and short repeated contacts between the same people. The latter scenario should have a larger probability than the former to become infected. While the distribution of $p_i[n]$ is homogeneous without a clear threshold for an infectious contact, $P_i[n]$ presents a very polarized distribution where the values are accumulated in the extremes: P_i = 0 or P_i = 1 (see Figure 10a). Thus, $P_i[n]$ is better indicator than $p_i[n]$ to separate infectious from non-infectious contacts. A contact is then considered infectious when this probability exceeds a certain threshold, $P_i[n] > p_c$.

The hyperparameters of the contact model (T, r, p_c) are obtained by calibrating the model using only the contacts between infected people to reproduce the basic reproduction number $R_0 = 2.78$ in Ceará in the month of March, 2020 (see Appendix B.2). We obtain T = 30 min, r = 8 m and p_c = 0.9. Thus, a contact is defined with probability one when exposure is at least 30 minutes within a distance $\ll 8$m. This calibration procedure provides the partial transmission tree of the outbreak from patient zero to the end of the calibration period shown in Figure 9c.

3.3 Transmission network model

Above we gave a probabilistic definition of a contact, i.e the probability that two persons (x and y) had a single contact its proportional to $p_1[n]$. In what follow we always focus on contacts between and infected persons and an healthy person, with the outcome being the chain transmission network with a tree-like structure. However, Eq. (3.5) is more general and can be applied in other studies that go beyond disease transmission. We first trace the trajectories of confirmed COVID-19 patients to search for contacts -14/+7 days from the onset of symptoms (see Appendix B.1). The interactions between patients and the healthy persons in the dataset identify the first layer $l = 1$ of contacts. Once in contact with a confirmed COVID-19 patients, the healthy person its likely ($\sim P_i[n]$) to be infected too. And therefore she/he can also spread the virus, defining in such a

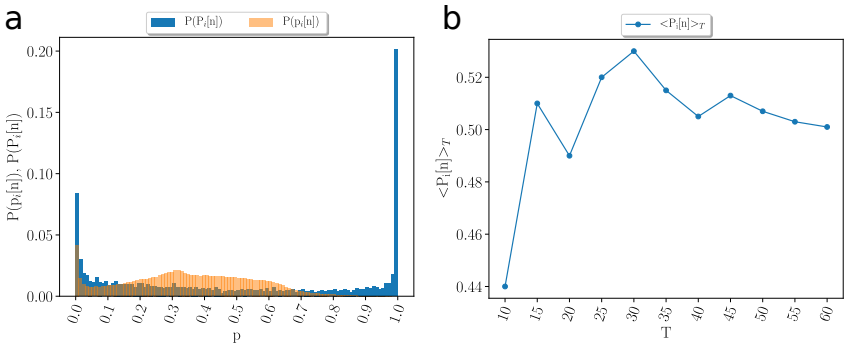


Figure 10: (a) Probability distribution of $p_i[n] = p_d[n] \cdot p_t[n]$ (orange) and the recursive form $P_i[n]$ defined in Eq. (3.5) (blue). The $P_i[n]$ are polarized to 0 and 1 becoming the best thresholded metric to use to consider a contact as infectious. (b) Average value $\langle P_i[n] \rangle_T$ as a function of the time window T of the spatio-temporal contact area. $P_i[n]$ has a peak at T = 30 min; it decreases for $T > 30$ min and increase for $T < 30$ min as a function of T. The decreasing behaviour is what is expected, thus, 30 min is the minimum bound for the correct value of T.

way a second layer of contacts. In general it is possibles to define an arbitrary number of layers, by considering the contacts between the people belonging to the layer $l - 1$ and all the other people which do not belong to any layer. The zero layer $l = 0$ consists of the contacts between the confirmed COVID-19 patients in the state of in Ceará. From the first contact layer, we add four layers of contacts to constitute the contact network of transmission that is used to monitor the progression of the pandemic. The time-varying network is aggregated to a snapshot defined over a time window of a week [18]. We find that other aggregation windows give similar results as presented. We analyze the spatio-temporal properties of such contact network (see Appendix B.3 for more details). The government of the State of Ceará imposed a mass quarantine on March 19, 2020 which led to a decrease in people's mobility by 56.5% as shown in Figure 11a. During the lockdown, only the displacements of essential workers were allowed. A large decrease in mobility is also observed across all Latin America, see [46].

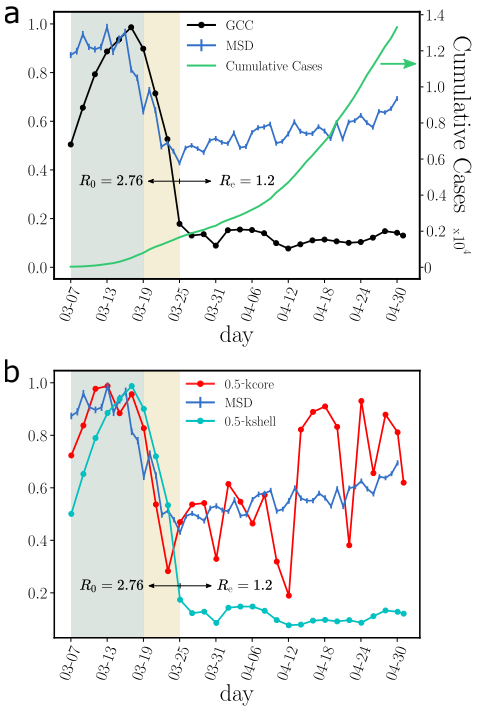


Figure 11: **(a)** Evolution for different metrics in Ceará, Brazil, previous to the mass quarantine (grey area), right after the imposed quarantine (yellow area) and later. The plot shows the root mean square displacement (MSD) normalized by the maximum value over the total period (blue), the cumulative number of cases (green) and the size of the GCC normalized by the maximum value over the total period (black). The uncertainty corresponds to the standard error (SE). The mobility data is showcased in the Grandata-United Nations Development Programme map shown in `https://covid.grandata.com`. The initial rise in GCC is due to the lack of data before March 1. **(b)** The plot shows the 0.5-kcore size (red), the 0.5-kshell size (cyan) all normalized by their respective maximum value pre-lockdown. While the size of the 0.5-kshell is reduced drastically during the lockdown, the 0.5-kcore was not reduced as much and keeps increasing, contributing to sustain the pandemic. The 0.5-kcore seems to follow the trend in the MSD, which we plot again to show this trend.

3.3.1 Giant connected component (GCC)

To understand the effect of the lockdown on the contact network, we think by analogy with a "bond percolation" process [120, 18, 44]. The percolation problem studies the dismantling of networks under removal of nodes or links (bond percolation), as well as associated problems of disease transmission [18, 44], robustness and resilience of networks [114, 115]. The global disruption in network connectivity is monitored by studying the normalized size of the giant connected component (GCC). The GCC is the largest connected subgraph of a graph, i.e., in the GCC there is always a pathway to reach a node from any other node. Formally, the network dismantling occurs at a critical percolation transition threshold q_c of the fraction of removed nodes q. The GCC occupies a fraction larger than zero for $q < q_c$ and vanishes otherwise. The vanishing of GCC at q_c marks the percolation transition between two phases, namely, a connected phase and a disconnected one.

One can remove nodes or links at random [120, 18] or by following optimized strategies to break the GCC with the minimal number of removals. We study different strategies to destroy the GCC with minimal removals based on removing nodes in the network by ranking them according to different centralities. All strategies are adaptive, meaning that the ranking is recalculated after every removal. We use:

- Degree strategy: We rank the nodes by their degree (number of contacts) from top (hubs) to low degree [114, 115] and then remove nodes starting from the hubs adaptively. We note that the process of removal of nodes from the network is a numerical trick commonly utilized to find the minimal number of hubs to dismantle the GCC. It is known in the literature that by removing the hubs adaptively, one is able to find a set of hubs that dismantle the network faster than by removing the hubs without adapting the network, see [114, 50, 116] for details. This is a purely algorithmic procedure to find the best set and the effect on the size of the GCC is the same as if one removes the nodes adaptively or at once. Thus, during the implementation of a quarantine in a real setting, the hubs

obtained from the algorithm should be quarantine at once as soon as the hubs can be contacted by the contact tracer app (see Section 3.5).

- K-core strategy: We rank the nodes by their occupancy in the k-shells of the network. The highest rank corresponds to the inner k-shell, that is the maximal k-core at $k_{\rm core}^{\rm max}$. Since each k-shell is formed by many nodes, we then rank the nodes inside a given k-shell by their degree. We then remove each k-shell in turn, always recalculating the ranking after every removal [50].

- Collective Influence (CI) strategy: we calculate the CI of each node according to Ref. [116] with $\ell = 3$ and remove the nodes with CI adaptive.

- Betweenness Centrality (BC) strategy: We rank the nodes by their BC and remove them one by one from highest to lowest, adaptively [52, 53, 121] The betweenness centrality of a node is proportional to the number of shortest paths that pass through the node [52, 53, 121]. It is calculated by considering all the pair of nodes in the network and calculating the shortest path between each pair. This methods was found in previous simulations to be a good predictor of a node's epidemic influence in a contact network [122]. For larger datasets used in this study, approximate fast algorithms can be used to calculated BC. See Ref. [123].

- We also use other strategies, like eigenvector-based centralities and combinations of other centralities to characterize the node importance [124].

We note that removing the nodes in an adaptive strategy is just a numerical trick to find a better set of spreaders. But the effect on the network is the same whether we remove the identified top spreaders one by one or all at once. Thus, in the implementation in a real quarantine, there is no need to contact the spreaders to quarantine one by one, but they should be notified all together that they should quarantine. Furthermore, after notification, we do not remove them from the analysis,

but we keep tracking them in case that they could infect new people in future networks (see Section 3.5).

We test the most efficient strategy to dismantle the transmission network. We plot the normalized size of the GCC, $G(q)$, that is, the number of nodes in the GCC after removal of a fraction of q nodes divided by the size of the GCC at $q = 0$. As nodes are removed from the network, we search for the strategy that provides the minimal removal with the maximal damage to the GCC. The best strategy over all the networks studied across all of the above ranking is the high BC strategy. The effect of a lockdown is to reduce the number of connections among people and therefore it acts as a percolation process. We monitor the GCC of the transmission network before and after the lockdown. We find a drastic percolation transition [120, 18, 18] within 6 days of the implementation of the lockdown on March 19, when the GCC is almost fully dismantled decreasing by 89.6% of its pre-lockdown size (Figure 11a). Despite the disintegration of the GCC, the cumulative number of cases kept growing albeit at a lower rate (Figure 11a). We find that the mass quarantine was able to reduce the basic reproduction number from $R_0 = 2.78$ before lockdown to an effective reproduction number of $R_e = 1.2$ after the lockdown (Figure 11a). Despite this disruption in the network connectivity, R_e has not decreased below one, as it would have been needed to curb the spread of the disease.

The drastic reduction in the GCC is visually apparent in the contact networks in Figure 12. Before lockdown on March 19 (Figure 12a), the network is a strongly-connected unstructured "hairball". Eight days into the lockdown on March 27 (Figure 12b), the network has been untangled into a set of strongly-connected modules integrated by tenuous paths of contacts. This structure is even more pronounced a few weeks later on April 28 (Figure 12c).

3.3.2 Superspreading k-core structures

The highly connected modules found in Figure 12b and 12c are k-core structures [47, 48, 49, 50] of higher complexity than the GCC (which is a

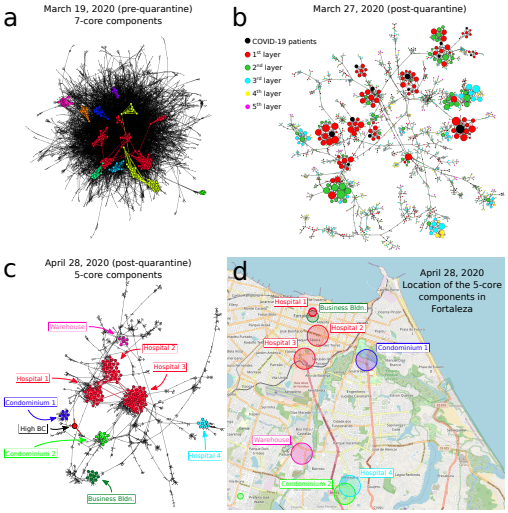


Figure 12: Disease transmission networks in the state of Ceará over time before and after the lockdown on March 19, 2020. **(a)** Transmission network on March 19 (pre-lockdown). A hairball highly-connected network is observed. The disconnected components of the 7-core ($k_{\text{core}}^{\max} = 12$ in this network) are colored. These components are well connected into the hairball network as expected since mobility and connectivity is high. **(b)** The pre-quarantine hairball in **(a)** has been untangled and the k-cores have emerged 8 days into the lockdown on March 27. Here, we color the nodes according to layers of the transmission network starting at COVID-19 patient (black nodes). Size of nodes is according degree. **(c)** Network on April 28 including the components of the 5-core in different colors ($k_{\text{core}}^{\max} = 7$ for this network). Visible is the high betweenness centrality node representing the weak-link of this k-core. **(d)** We plot the location of the contacts in the map of Fortaleza constituting the components of the 5-core of the April 28 in **(c)**. The size of the circles in the map corresponds to the number of contacts inside each location. The colors correspond to the clusters of the 5-core in **(c)**. The 5-core sustaining transmission is composed of clusters of contacts localized in hospitals, large warehouses and business buildings. Hospital 3, one of the largest in Fortaleza, constitutes the maximal $k_{\text{core}}^{\max} = 7$ of the pandemic.

1-core), that are known to sustain an outbreak even when the GCC has been disintegrated [18, 50].

The k-core of a graph is the maximal subgraph made of nodes with

degree k or more [47, 48, 49, 50]. A k-shell of a graph is composed by all the nodes that belong to the k-core but not to the (k+1)-core. See Figures 13 and 14 for definitions in a network with 3-shells, i.e., with a maximal 3-core ($k_{\rm core}^{\rm max} = 3$), and Figure 15 for an example in a real network.

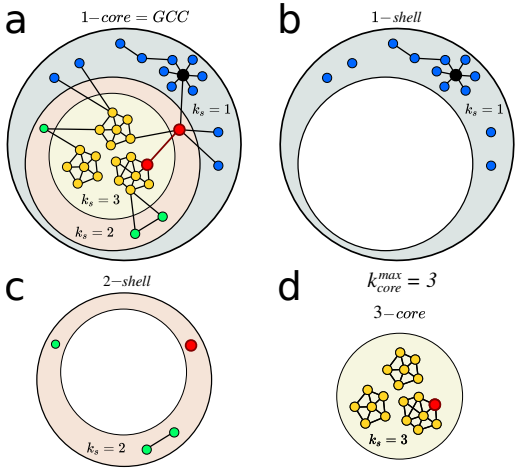


Figure 13: (a) A sample network with 3 shells. The k-shell index k_s is not necessarily associated with other centralities. Here, the hub of the network in black with $k = 7$ is in the 1-shell, $k_s = 1$. The two top node in betweenness centrality, highlighted in red, belong to the 2-shell and the 3 shell, respectively. The 1-core is equivalent to the GCC. **(b)** The nodes with $k_s = 1$ form the 1-shell, **(c)** the nodes with $k_s = 2$ form the 2-shell, and **(d)** the nodes with $k_s = 3$ form the 3-shell which is also the 3-core.

The k-shell decomposition assigns each node to a k-shell in the network. The k-cores are nested structures and k-shells are disjointed; e.g., the 2-core contains the 3-core and so on, and the 2-core is formed by the 3-core plus the 2-shell. By definition, the GCC is the largest 1-core. The maximal k-core is the inner subgraph of the network and it is indexed by $k_{\rm core}^{\rm max}$ index. The low k-shell are the peripheric shells.

In practice, the k-core of a network is obtained by iteratively removing all nodes with degree smaller than k. One starts the removal process by removing all nodes of degree one, Figure 14. After the first removal, nodes that initially had degree larger than one may end up with degree

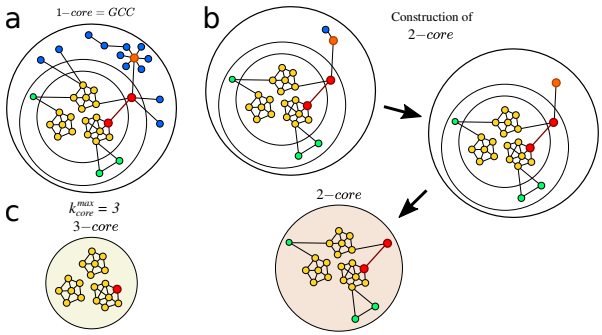


Figure 14: (a) We start the k-shell decomposition with a network configuration where every node has at least degree $k = 1$. This set of nodes forms a 1-core. **(b)** Then, every node with $k = 1$ is iteratively removed to obtain the 2-core. As one can see, the removal of these nodes changes the degree distribution. Thus, nodes are removed until all remaining nodes are left with $k \geq 2$. **(c)** Following the k-shell decomposition nodes are removed until we obtain the 3-core. The 3-core can be made of multiple disconnected clusters.

equal to one. Then, one repeats the process until no remaining node in the network has degree equal to one, or equivalently, every node in the network has degree at least equal to 2. This set of nodes with $k = 2$, or higher, is the 2-core. The other k-cores are obtained in analogous manner.

It is important to note that a k-core can be composed of disconnected clusters or components. For instance, the example 3-core in Figure 14c contains 3 components. These three components are disconnected in the 3-core, but they are integrated in the network by nodes belonging to the 2-shell as shown in Figure 14a. This is an important property of the k-cores found in the transmission networks during the lockdowns. For instance, the network of Ceará from March 27 plotted in Figure 12b has a rich k-core structure shown in Figure 15 where, for instance, the 6-core is composed of 5 disconnected components. This is an important property for an strategy based on betweenness centrality. Typically, we find that these components are joined together by nodes in lower k-shells, which are identified by their high betweenness centrality. Then, the k-cores components can be relatively easily dismantled by a few removals outside the k-cores. The maximal k-cores are composed of nodes

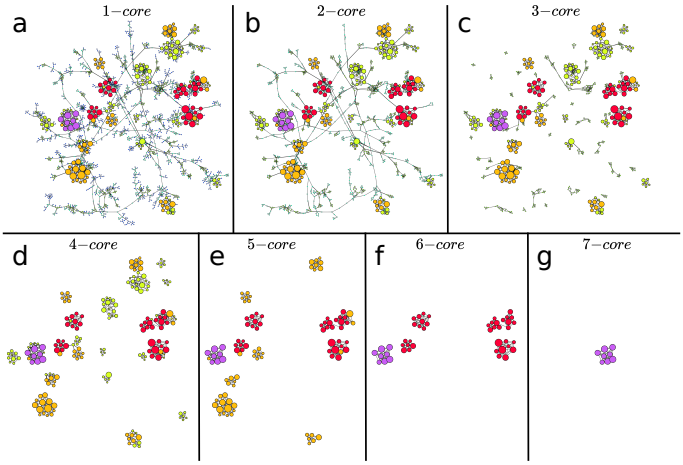


Figure 15: Example of k-core and k-shell structure in the network plotted in Figure 12b obtained during the lockdown. Here the colors are set by the k-shell occupancy of each node. Each k-core is composed by the k-shell plus the (k+1)-core. The k-cores are nested structures. For instance, the 5-core in **(e)** is composed by the 5-shell (yellow nodes) and the 6-core, which, in turn, is composed by the 6-shell (in red) and the 7-core (in purple). Since the 7-core is the maximal k-core, $k_{\rm core}^{\rm max} = 7$ for this network, then the 7-core is also the 7-shell. In this network the 0.5-kcore is the 4-kcore and the 0.5-kshell is composed by the 1-shell plus the 2-shell and the 3-shell. We notice how a given k-core can be composed of many disconnected components. For instance, the 6-core is composed by 5 disconnected components. This is important, since each component of a given k-core can be localized in different areas, like different hospitals, in the map, see for instance, Figure 12c and 12d. It is also visually apparent that to destroy this network, a direct 'attack' to the high k-cores is not optimal. Instead, removing the high BC nodes that populate the lower k-shells is the best strategy. We plot each k-core in turn: **(a)** 1-core, **(b)** 2-core, **(c)** 3-core, **(d)** 4-core, **(e)** 5-core, **(f)** 6-core and **(g)** 7-core.

with high degree that connect with other nodes of high degree, which in turn also connect with others of high degree, and so forth. This implies that the k-cores do not have dangling ends made of nodes with degree smaller than k. That is, the k-cores are close, in a sense. The k-shell decomposition then cleans the network of those low degree dangling ends

in a systematic way and reveals the core of the network, which is the most important part for spreading [125]. Notice that a hub can be in the maximal k-core or in an outer k-shell according to how the hub is connected. For instance, the red hub in Figure 14a is in the 3-core because it is also connected to other hubs with 3 or more connections. However, the orange hub is in the 1-shell in Figure 14a because it is connected with nodes with low degree.

Monitoring the k-cores of the networks as a function of time we find that before the lockdown the maximal k-core index is in average around $k_{\text{core}}^{\max} \approx 12$ and then drops to half of this value with a maximal 6-core in average during the lockdown ($k_{\text{core}}^{\max} \approx 6$). Figure 16 shows this drop in the maximum k-core index from $k_{\text{core}}^{\max} \approx 12$ to $k_{\text{core}}^{\max} \approx 6$, in average. Since $k_{\text{core}}^{\max}$ changes with time, to study the occupancies of the k-cores and k-shells across the quarantine transition, we define the $\epsilon-$kcore as the k-core with k such that $k = \lceil \epsilon\, k_{\text{core}}^{\max} \rceil$. The complement of the ϵ-kcore is the ϵ-kshell defined as the union of the remaining k-shells with k such that $k = 1, 2, ..., \lceil \epsilon\, k_{\text{core}}^{\max} \rceil - 1$. Thus, the union of the ϵ-kcore and the ϵ-kshell constitute all the network. In the paper we consider $\epsilon = 0.5$ that divides the k-shells in two.

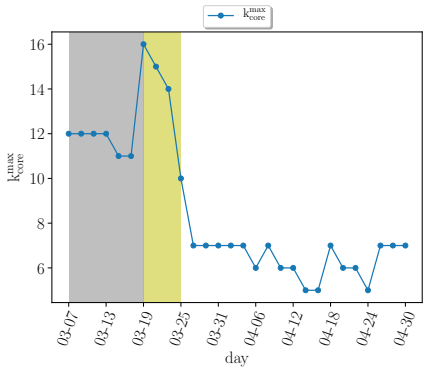


Figure 16: Evolution of maximum k-core index $k_{\text{core}}^{\max}$ versus time previous to the quarantine (grey area), right after the quarantine (yellow area) and later. We see how the maximum k-core index drops drastically after the mass quarantine.

The high k-cores are known from network science studies to be the reservoir of disease transmission persistence [18, 50]. On the contrary, low peripheral k-shells do not contribute as much to the spread as the high inner k-cores.

Figure 11b shows that despite the disappearance of the GCC, there is a significant maximal k-core that was not dismantled by the mass quarantine. The figure shows that the outer k-shells of the transmission network (i.e., the 0.5-kshell defined as the union of the k-shells with $k = 1, 2, ..., \lceil 1/2\, k_{\rm core}^{\rm max} \rceil - 1$, see Appendix B) are disintegrated in the lockdown, decreasing by 91% with respect to their pre-quarantine size, in tandem with the GCC. However, the inner k-core (i.e., the 0.5-kcore defined as the k-core with $k = \lceil 1/2\, k_{\rm core}^{\rm max} \rceil$, see Appendix B) persists in the lockdown. The figure shows that the decrease of the 0.5-kcore is only 50% compared to the 91% decrease of the 0.5-kshell; the former even increases slightly at the end of April, following the same trend in mobility (see Figure 11b). This process is visually corroborated in the evolution of the networks seen from Figure 12a to 12c where we observe the disappearance of the peripheral k-shells and the persistence of the maximal k-core. Indeed, the unessential contacts in the peripheral k-shells may have been first pruned during social distancing.

Using numerical simulations, we corroborate that the infection can persist in these high k-cores of the network while virus persistence in outer k-shells is less important [18, 50]. We use a SIR model on the transmission network (Figure 17a and Figure 18a) showing that the maximal k-cores of the network sustain the spreading of the disease more efficiently than the outer k-shells. Thus, the maximal k-core components of the contact network are plausible drivers of disease transmission. Apart from this structural explanation (i.e., k-core), epidemiological factors may also play a role in the persistence of the disease, such as a transition of the disease to vulnerable communities with high demographic density, or with large inhabitants per household where isolation is poorly fulfilled.

When we plot the geolocation of the contacts forming the maximal k-core in the map of Ceará, we find that these contacts take place in highly

transited areas of the capital Fortaleza, such as hospitals, business buildings, warehouses as well as large condominiums, see Figure 12d. These contacts generate superspreading k-core events that generalize the conventional notion of superspreaders, which refer mainly to individuals with large number of transmission contacts [126, 127, 128]. However, connections are not everything [114, 115]. K-core superspreaders not only generate a large number of transmission contacts, but their contacts are also highly connected people, and so forth.

3.4 Results

The existence of k-cores in the transmission network suggests that a more structured quarantine could be deployed to either isolate or destroy those cores that help maintain the spread of the virus. We perform an optimal percolation analysis [114, 115, 116] to find the minimal number of people necessary to quarantine that will dismantle the transmission network. We follow different strategies to find the optimal breakdown of the network by ranking the nodes based on (1) the number of contacts (hub-removal) [114, 115, 18], (2) the largest k-shells and then by the degree inside the k-shells [18, 50], (3) the collective influence algorithm for optimal percolation [116], and (4) betweenness centrality [52, 53, 121, 122].

Figure 17b shows the normalized size of the GCC versus the fraction of removal nodes following different strategies, as well as a random null model of removal in a typical network under lockdown in April 28 (March 19 pre-lockdown results are plotted in SI Figure 18b). While the disease can persist in the k-cores (Figure 17a), quarantining people directly inside the maximal k-core is not an optimal strategy. The reason is that k-cores are populated by hyper-connected hubs that require many removals to break the GCC [122] (around 7%, see Figure 17b). For the same reason, removing directly the hubs is not the optimal strategy either, since the hubs are within the maximal k-core and not outside. A collective influence strategy [116] improves over hub-removal since it takes into account how hubs are spatially distributed, yet, it is far from optimal. Clearly, Figure 17b shows that the best strategy is to quarantine

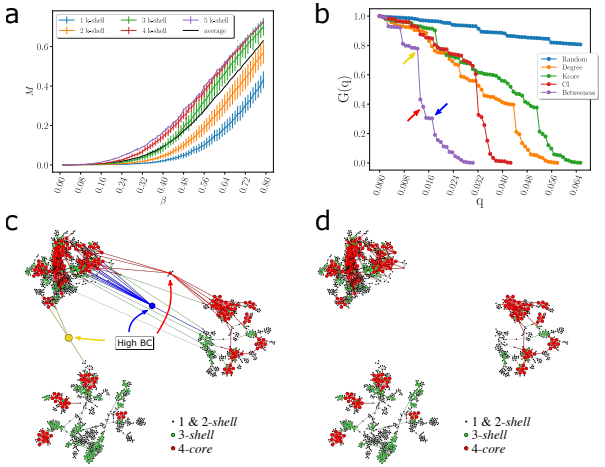


Figure 17: (a) Average size of infected population, M [50], in an outbreak average over all starting nodes in a k-shell as a function of the probability of infection β for a SIR model on the network in Figure 12c during the lockdown. The black is the average value over all the network. The average divides the k-shell contribution to the spreading of the virus in two groups: above and below the average. The 0.5-kcores have maximal spreading and the 0.5-kshell have minimal spreading. Error bars correspond to a confidence interval of 95%. **(b)** Optimal percolation analysis performed over the network in Figure 12c during the lockdown in following different attack strategies and their effect on the size of the largest connected component $G(q)$ versus the removal node fraction, q. Nodes are removed (in order of increasing efficiency): randomly (blue); by the highest k-shell followed by high degree inside the k-shell [50]; by highest degree (orange); by collective influence (red) [116]; and by the highest value of betweenness centrality (green) [52, 53]. After each removal we re-compute all metrics. The most optimal strategy among those studied is removing the nodes by the highest value of betweenness centrality. **(c)-(d)** Effect of removing three high betweenness centrality nodes shown in Figure 17b in the network of Figure 12c. **(c)** We show the 2-core component of the network after the removal of 12 high betweenness centrality nodes. The red node is the one with the highest betweenness centrality value (next node to remove, 13th) and the blue node is the 14th removal. Different k-cores and k-shell are in different colors. **(d)** Network k-cores are disintegrated after the removal of the high BC nodes.

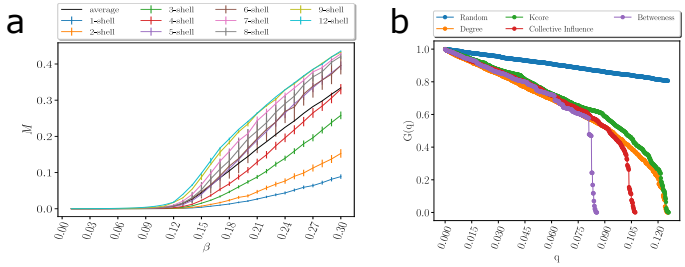


Figure 18: (a) Amount of infected population ($M = \sum \frac{M_i}{N}$ see [50]) when the spreading starts in a given node in a k-shell as a function of the probability of infection β for a SIR model on the same network on March 19 in Fig. 12a. in pre-quarantine Ceará. The black is the average value over all the starting nodes in the network. The average divides the shell contribution to the spreading of the virus in two groups above and below the average. The 0.5-kcore composed of the 6-core ($k_{\rm core}^{\rm max} = 12$ in this network) which contains nodes from the 6-shell to the 12-shell, has maximal spreading. the 0.5-kshell which is composed by the remaining shell from 1-shell to 5-shell has minimal spreading, below the average. **(b)** Optimal percolation analysis performed over the network in Fig. 12a before the quarantine on March 19 in Ceará with different attack strategies and their effect on the size of the largest connected component $G(q)$ versus the removal node fraction, q. Depending on the strategy nodes are removed: randomly (blue), by the highest value of betweenness centrality (green) [52, 53], degree (orange), collective influence (red) [116], and by the highest k-shell followed by high degree inside the k-shell [50]. After each removal we re-compute all the metrics. The best strategy among those studied is removing the nodes directly by the highest value of betweenness centrality.

people by their betweenness centrality. By removing just the top 1.6-2% of the high betweenness centrality people, the GCC is disintegrated. This is consistent with the particular structure of the transmission networks seen in Figure 12b, c and Figure 17.

The betweenness centrality of a node is proportional to the number of shortest paths in the network going through that node. Thus, given the particular structure of the networks in Figures 12b, c, and Figure 17c, the high betweenness centrality nodes are the bottlenecks of the network, i.e., loosely-connected bridges between the largely-connected k-cores components. These connectors are the celebrated 'weak links', fun-

damental concept in sociology proposed by Granovetter [51], according to which, strong ties (i.e., contacts in the k-cores) clump together forming clusters. A strategically located weak tie between these densely 'knit clumps', then becomes the crucial bridge that transmits the disease (or information [51]) between k-cores.

These weak links are people traveling among the different k-cores components allowing the disease to escape the cores into the rest of society. These bridges are displayed in the network of Figure 17c as yellow, blue and red nodes. The removal of these high betweenness centrality people disconnects the k-core components of the network entirely, as shown in Figure 17d, halting the disease transmission from one core to the other [129, 122].

An important finding is that quarantining the large superspreading k-cores is neither optimal (as shown in Figure 17b, green curve) nor practical, since they are mainly comprised by chiefly essential workers who need to remain operational (Figure 12d). Thus, the best strategy, in conjunction with the mass quarantine, is then to disconnect these k-cores from the rest of the social network (Figures 17c and 17d), rather than quarantining the people inside the k-cores. This can be performed by quarantining the high betweenness centrality weak-links that simultaneously preserve the operational k-cores. However, individuals belonging to the maximal k-cores should be tested at a higher frequency to promptly detect their infectiousness before the symptoms start, to help control the spreading inside the k-cores.

3.5 Discussion

Isolating the k-core structures by quarantining the high betweenness centrality weak links proves to be the most effective way to dismantle the GCC of the disease while keeping essential k-cores working. While destroying the strong links and cores is a less manageable task to execute and control, isolating the weak links between cores is a more feasible task that will assure the dismantling of the GCC. In other words, if one core is infected, the disease will be controlled within that core and not extended

to the rest of society.

The algorithms we propose, which should help policy makers in fighting the current pandemic, are implemented following the guidelines of the WHO for developing contact tracing programs using traditional or digital technologies. In [130] the World Health Organization, specifically recommends a daily monitoring of infected people, daily tracing of their contacts and daily application of quarantines based on the identified contacts.

Thus, our protocol could be implemented by local authorities by the use of a mobile app, which would (i) first identify the infected people day by day, (ii) then build the contact network back in time starting from those infections, (iii) compute if a user is a potential spreader, and (iv) if so, notify the user to quarantine, after which the user keeps being monitored.

The intervention strategy can be implemented as follows:

1. Monitor daily the number of infected people in the database. Every day the database is updated with the new cases including the information on the patient's first date of symptoms.

2. Starting from the day of first symptoms of each patient, monitor back in time and in future days the number of contacts using digital contact tracing. From the identified first layer of contacts, obtain the chain of transmission by constructing the possible contact network up to a given number of layers. This is the transmission network at day t. This network takes into account possible pathways of the spreading.

3. After identifying all members of the transmission network, notify them of possible exposures. Additionally, perform network techniques to identify the people to dismantle the GCC and further inform them to quarantine, isolate, test or vaccinate. Optimization is applied numerically at this stage, by following the analysis in Figure 17 over different dismantling strategies. Send a message via mobile app to isolate to everyone in the transmission chain, in particular to those in the first layer of contact. Send also a message

via mobile app to the obtained list of top spreaders to further suggest to quarantine. The quarantined people should not be removed from the analysis in the subsequent networks at time $t+1$ since the quarantined person could still create new contacts in the future, if the quarantine is not followed strictly. The quarantined people are not removed from the analysis and their trajectories are being monitored in future times to detect future contacts.

4. At day $t+1$ repeat the same procedure from Point 1-3. Add to the database the new number of reported cases at day $t+1$ and build the new contact network explained in Point 2.

5. Repeat the above procedure daily for maximum effectivity.

The core of our algorithm is the capability to detect possible infected persons. As such, it can helps the governments around the world in order to curb the spread of coronavirus [104, 105, 106, 107, 108, 109, 110, 111]. Such identification can guide to targeted vaccinations/quarantine and it is of primary importance in the second phase of reopening economies across the world and, in particular, in developing countries where resources are scarce. In the Appendix B we addresses issues of sampling bias and coverage of the data relevant for implementation. Overall, our network-based optimized protocol is reproducible in any setting and could become an efficient solution to halt the critical progress of the COVID-19 pandemic worldwide drawing upon effective quarantines with minimal disruptions.

Chapter 4

The social listener: an AI alternative to traditional polls

The results presented in this Chapter are based on the work [54], currently under submission.

4.1 Introduction

Traditional polling methods [131] using random digit dial phone interviews, opt-in samples of online surveys, and interactive voice response are failing to predict election outcomes across the world [132, 133]. The failure of traditional surveys has also been widely discussed in the press [134] and on specialized literature. For instance, the victory of Donald Trump in the US 2016 presidential election came as a shock to many, as none of the pollsters and political journalists, including those in Trump's campaign, could predict this victory [135].

One of the reasons of this failure is that the percentage of response to traditionally conducted surveys has decreased and it is becoming increasingly difficult to get people's opinion [136, 137]. Response rates in telephone polls with live interviewers continue to decline, as it has

reached 6% lower limit recently [137]. Response rates could be even lower for other methodologies, like internet polling or interactive voice response. There is increasing evidence [136, 137] that the nonresponse bias might be the reason that polls are not producing accurately matched election results. However, this is not the only problem of traditional methods of polling. Live interviews are affected by social desirability biases [138], i.e. the tendency of research subjects to give socially desirable responses instead of choosing responses that are reflective of their true feelings. Moreover, polls are not able to detect sudden change of opinion due to some particular events or circumstances, since the process of opinions collection is time consuming. All these peculiarities together make impossible for traditional polls to correctly predict the results of the elections.

Monitoring social networks represents an alternative for capturing people's opinions since it overcomes the low-response rate problem and it is less susceptible to social desirability biases [138]. Indeed, social media users continuously express their political preferences in online discussion without being exposed to direct questions. One of the most studied social networks is the microblogging platform Twitter [139, 140, 141, 28]. Twitter's based work generally consist of three main steps: data collection, data processing and data analysis. The collection of the tweets is often based on a public API of Twitter. It is a common practice to collect tweets by filtering according to specific queries, as for example the name of the candidates in the case of elections. Data processing includes all those techniques which aim to guarantee the credibility of the Twitter dataset. This is, for example, bots detection and spam removal [28]. Data analysis, the core of all these studies, can be simplified in three main approaches: volume analysis, sentiment analysis and network analysis.

Scholars used the number of mentions for a party of a candidate in order to forecast the result of the 2009 German parliament election [142]. While their technique has attracted many criticisms [143], their work was of inspiration to many other researchers. Gaurav *et al.* [144] proposed a model, based on the number of times the name of a candidate is mentioned in tweets prior to elections, to predict the winner of three presi-

dential elections held in Latin America (Venezuela, Paraguay, Ecuador) from February to April, 2013. Based on volumetric analyses are also the works of Lui *et al.* [145] and Bermingham [146]. Ceron *et al.* [147] performed a sentiment analysis study on the tweets to check the popularity of political candidates in the Italian parliamentary election of 2011 and in the French presidential election of 2012. Caldarelli *et al.* [148] used the derivative of the volume to forecast the results of Italian elections. Singh *et al.* [149] employed sentiment analysis to predict victory of Trump in the election of 2016. The same author proposed a method [150] based on sentiment analyses and machine learning on historical data to predict the number of seats that contesting parties were likely to win in the Punjab election of 2017. Other works [151, 152] used social networks analyses in order to identify the position of a party in the online community by measuring its centrality. The most supported parties are in general those with an higher centrality.

Despite the large amount of literature, the debate about whether Twitter can be used to infer political opinions is still open. Online social networks are continuously filled by false, erroneous data through trolls, bots and misinformation campaigns to a level that distinguish between what is genuine and what is not is in general difficult. By virtue of this, the great challenge of algorithms and AI is to discover and interpret real data from 'junk data' that could lead to accurate predictions of electoral or opinion trends. Beside, it should be noted that the opinions of Twitter users may not be representative of the entire population [153, 139].

In this work we first investigate in section 4.2 why the traditional polls fail to predict the results of the primary presidential election in Argentina on August 2019. This is possible thanks to the exclusive access to the raw data of the polling conducted by one of the most reliable pollsters in Argentina, Elypsis. We find that a poor demographic representation combined with the inconsistency of opinion' respondents before and after the elections are the main reason why polls fail. To overcame these problems, we propose an AI model to predict electors trend of opinions in social media in section 4.3. By using machine learning we uncover political and electoral trends without directly asking people what they

think, but trying to predict and interpret the enormous amount of data they produce in online social media [141, 28, 154, 155]. Big data analysis overcomes the low response rate problem. By re-weighting the Twitter populations to the Census data, we match the distribution of the population' statistics (age, gender) and the statistics of the real population (given by the Census Bureau) [153]. The real time data processing which underlies our AI algorithm allows us to detect sudden change of opinions, and therefore different loyalty classes towards each candidate. We will show that a cumulative analysis performed on the loyalties classes considerably improves the results of [141], based on instantaneous predictions. Instantaneous predictions, as well as pollsters predictions, are subject to high fluctuations which undermine the reliability of the prediction itself. We validate the algorithm on the general election in Argentina. Our results (section 4.4) show that AI can capture the public opinion more precisely and more efficiently than traditional polls. We summarize the result in the discussion, section 4.5.

4.2 Why are pollsters failing to predict elections?

The events leading up to the recent primary election in Argentina are a telling example of the failure of the polling industry [156, 157, 158]. On the primary election day on August 11, 2019 (called PASO in Spanish: Primarias, Abiertas, Simultáneas y Obligatorias; in English: Open, Simultaneous, and Obligatory Primaries), none of the pollsters in the country predicted the wide 16% margin of presidential candidate Alberto Fernández (AF) over the president Mauricio Macri (MM). Primaries in Argentina are obligatory, happening for all political parties at the same time, and the two main parties presented only one candidate each, thus transforming the primaries into a de-facto presidential contest.

Figure 19 shows the comparison between the official results (in red), our prediction (in blue, Model 3 explained below) and the polling average, computed as the average of the top five most trusted pollsters in Argentina [157, 158], i.e. Real Time Data, Management & Fit, Opinaia, Giacobbe and Elypsis (in green). Macri was clearly defeated by Fernan-

dez by +16%, a result captured by our predictions. While the average pollster predicted Fernandez with a slight advantage in the primary, the estimated percentage of each candidates were, in general, really close reaching in some occasion a difference of just one percentage point [159]. Elypsis in particular predicted that Macri would win for one percentage point [160]. This virtual tie predicted by the pollsters was largely considered to be a win for the incumbent candidate Macri since he was supposed to gain all the votes left by the third party options in the subsequent presidential election and eventually win the election in a runoff.

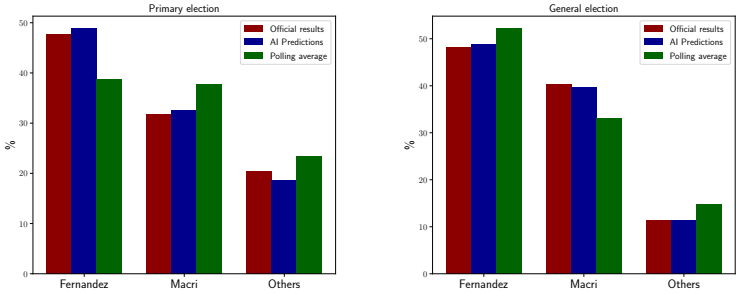


Figure 19: Comparison between polling average (green), official results (red) and our prediction (blue) for both the Primary (on the left: 2019-08-11) and the General election (on the right: 2017-10-27).

It is worth to stress at this point that Macri (right-leaning candidate) made of the internationalization of the market one of the main points of his campaign (favoring foreign investors and pro-business) while the left-leaning candidate Fernándezwas was supporting a national market. As a result of the predictions supported by all Argentinian pollsters giving Macri as the winner, the bond market rose excessively in the days preceding the primary election. The subsequent defeat of Macri by 16 percentage points at the primaries leads to a historic collapse of the MERVAL index by 40%, the bond market collapsed and some banks lost 1 billion dollars in the bet overnight [161, 162].

The failure of traditional polls is not associated with the impossibility of giving the exact right percentage for the candidates, but rather with the impossibility of predicting the enormous gap between them.

The Argentina primaries elections are not the only example of pollsters' failure. Unpredictable results seems to be associated whenever one of the candidates is a controversial figure in the political scenario. In Argentina, the eventual vice-presidential candidate accompanying Alberto Fernández was previous Argentinian president Cristina Fernández de Kirchner (CFK), who had faced corruption allegations and judicial processes and many Argentinians and the traditional media viewed as a controversial and divisive figure in Argentina politics and who is usually vilified by the traditional media. Notorious examples are Trump in the American presidential election of 2016 or Bolsonaro in the Brazilian general election of 2018. In the case of the Argentina primaries, the pollster failure led to economic disruption of the country at a national/international level [161].

Below we analyze the raw data of one of the most reliable polls, Elypsis (trusted specially by the president Macri and international investors [161, 162, 158, 159]) which as all the pollsters failed to predict the large gap between the two candidates for the primaries elections.

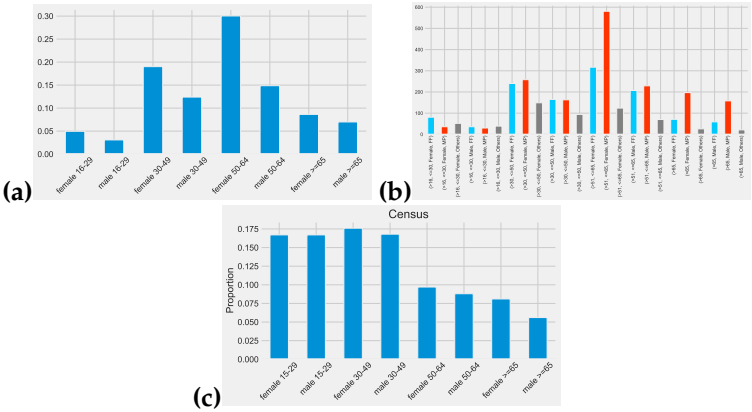


Figure 20: **(a)** Elypsis demographics before PASO. **(b)** Elypsis polling results before PASO separated by age and gender (Macri=red, Fernández= blue, Others=grey). Results are highly biases to older than 50 as compared with Census distribution. **(c)** Argentina Census Bureau 2010 demographic distribution by age and gender.

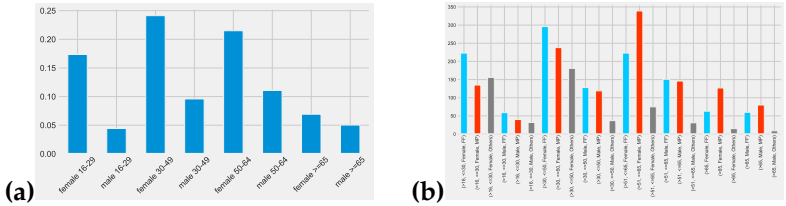


Figure 21: (a) Elypsis demographics after PASO. **(b)** Elypsis polling results after PASO separated by age and gender (Macri=red, Fernández= blue, Others=grey).

Figure 20a shows the age-gender distribution of the respondents to the survey conducted by Elypsis immediately before the PASO elections. Elypsis employs a combination of IVR of landline numbers complemented with online opt-in samples from Facebook. The vast majority of these online panels in Argentina, as well as in US, are made up of volunteers who were recruited online and who received some form of compensation for completing surveys, such as small amounts of money or frequent flyer miles. Figure 20b shows the number of respondents for Fernández (light blue), Macri (red) and Third Party (grey) grouped by age and gender. The Elypsis sample is peaked around the 50 years old group. This is strikingly different from the national population statistics obtained from the Argentinian Census Bureau shown in Figure 20c.

Elypsis data does not have significant coverage among people younger than 30, even though it has been conducted in Facebook. It shows a heavier tail on the right for older groups, while the national population (Census Bureau) has a less pronounced peak around the group of 30 years old and an heavier tail on the left for younger groups. The largest sampled group surveyed by Elypsis are females between 51 and 65 years old who are overwhelming in favor of Macri. In fact, in all groups above 30 years old, Macri is the clear favorite in the Elypsis poll. On the contrary, Twitter represents better the younger generations. It is important to consider again that the vote is obligatory in Argentina and it is permitted above 16 years, and the turnout of the youngest is quite substantial, thus, any pollster that does not capture their preferences is, in practice, doomed to

fail.

To deal with the mis-representation problem, pollsters adjust their raw results to population benchmarks distributions given by the Census Bureaus [136, 131] by weighting the raw data (sample-balancing or raking [163, 164]). The poll sample is weighted so it matches the population on a set of relevant demographic or political variables, for instance, age, gender, location and other socio-economic variables, like education level or income. Studies of the effectiveness of various weighting schemes suggest they reduce some (30 to 60%) of the error introduced by the biased sample, see [131]. However, when the raw data distribution is drastically under/over sampled as the Elypsis case, a small error in the most representative groups would propagate to produce inaccurate result.

As discussed above, the mis-representation is not the only problem which traditional pollsters methods face. Next, we analyze the longitudinal data taken on the same 1,900 respondents by Elypsis before and after the elections to investigate the social-desirability bias. We start with Figure 21a showing the Elypsis respondent distributions after PASO (notice that these respondents are different from the previous ones, and this is the reason why the age distribution change respect to the previous figure). By comparing Figure 20a before PASO with Figure 21a after PASO we first notice a change in the voters distributions. Younger groups are better represented after the election when compared to Figure 20a, although the data are still highly biased towards older generations. This implies that younger groups were, at least, more prone to answer the polls after the election than before.

Surprisingly, the female group with ages between 30 and 50 years voted for Fernández as indicated after the PASO polls, while before the PASO they responded mainly in favor of Macri. The male group of the same age shows a similar behavior, even if less pronounced. Let us notice that, according to Figure 20b, the groups of females/males between 30 and 50 years old are the most represented in the Census data and therefore may have an higher impact on the final result. These results can only be explained by admitting that voters did not say the true.

This is further corroborated by this unique longitudinal panel, as seen

in Table 2, revealing that people lied and hid their true voting intentions to the pollsters before the elections.

Who will you vote?	**Who did you vote?**			**Was Pre PASO vote true?**	
	AF-CFK	MM-MP	Other	Yes	No
AF-CFK	91%	2%	8%	91%	9%
MM-MP	6%	83%	11%	83%	17%
Lavagna	19%	9%	72%	56%	44%
Del Cano	25%	0%	75%	54%	46%
Espert	19%	14%	67%	53%	47%
Gomez Centurion	10%	8%	83%	69%	31%
Blank or Null	23%	4%	73%	47%	53%
Unknown or Others	53%	11%	36%		

Table 2: Vote disclosure analysis: "Who are you going to vote in the PASO" with "Who did you vote in the PASO" - using the same sampling and post-stratification methodology than in the Pre-PASO survey [165].

More specifically, when comparing "Who are you going to vote in the PASO" with "Who did you vote in the PASO" - using the same sampling and postratification methodology than in the Pre-PASO survey - it is found that about 18% of the people did not disclose their true vote, and the hidden vote was not unbiased.

- 91% of those who said "I will vote for Fernandez" did so, but only 83% in the case of Macri, who lost 6% to AF.
- "Secondary candidates" voters were much more volatile. Only 56% of those who said that they were going to vote for (third candidate) Lavagna disclosed their true vote, and 54%, 53% and 69% in the case of the candidates Del Caño, Espert and Gomez Centurion respectively.
- Alberto Fernández got almost 19% of the votes of those who chose a secondary candidate in the PASO Poll, and Mauricio Macri only 9%.
- Alberto Fernández received 46% of the votes of those who answered "Blank, Null or Unknow" before the PASO.

But, who hid - or not disclosed - their real vote? We find no significant difference between men and women or between education levels but we see a clear pattern in age demographics. 33% of those between 16 and 30 years changed their vote vs. their Pre PASO answer and only 13%, 10% and 14% on those between 31 and 50, 51 and 65 and more than 65, see Table 3.

	Revealed	**Not Revealed**
Man	83%	17%
Woman	81%	19%
Between 16 and 30	67%	33%
Between 31 and 50	87%	13%
Between 51 and 65	90%	10%
More than 65	89%	11%
Full Secondary	81%	19%
Incomplete Secondary	81%	19%
Full or incomplete Univ.	86%	14%
Total	82%	18%

Table 3: Hidden vote by demographics [165].

What did those who did not disclose their vote think about the candidates? Where they "closeted Kirchnerist" (party of AF and CFK) or did they bridge the gap between Macri-Fernández?

"Regular" images of Cristina Fernández, Macri and Alberto Fernández were much lower among those that did reveal their vote than among those who did not, see Table 4. Those who hid their votes look more nonpolarized, with a "Regular" image - No positive nor negative - of 21% on average, vs 6% / 10% of those who revealed the vote. CFK's negative image is higher than MM (48% vs 39%) in "non-revealed" and the opposite hold in the revealed (43% vs 50%). 35% of the "non-revealed" did not have (or hid) their opinion of Alberto Fernandez vs. 8% in the revealed.

This combined information shed light on PASO results and Polls consensus miss. In the PASO, AF was able to catch votes from all the candidates, and seduce voters from within the gap, "moderate" voters who

had a negative image of CFK and MM. He succeeded in standing himself as the "third candidate" bridging the gap, something that was not being fully captured by the polls, or that was decided at the last minute. This feature is most striking in young people, who may both have more "volatile" opinions and less prone to reveal them on traditional polls. This hidden-vote factor can explain by itself as much as 10% difference between "ex-ante" forecast and real results. Thus, standard polls methods failure may not have been related only to a bias in the sampling but, in the extraction of "True" information from surveyed people.

Image % of the total	**Revealed**				**Not Revealed**			
	Positve	Negative	Regular	NS/NC	Positve	Negative	Regular	NS/NC
CFK	45%	43%	6%	5%	20%	48%	22%	11%
MM	36%	50%	10%	4%	26%	39%	21%	14%
AF	45%	39%	7%	8%	15%	28%	28%	35%

Table 4: CFK, AF, and MM Image as % of the total [165].

Understanding why people lie is not the topic of this work even if, according to the literature the reasons could be many and related to desirability-bias. On one hand, participants may typically rush through the surveys to obtain their rewards and don't respond thoughtfully [136]. On the other hand, social-desirability bias [138, 166], i.e. the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others [166, 136] is another reason for people to hide their preference for controversial candidates like CFK, which leads to biased results.

We showed that low response rate, mis-representation and the social desirability bias/lies (which in the case of Elypsis biased more the younger representative) undermined the predictions on the Argentinian primary elections. Below we search for a suitable replacement using sampling methods for the modern era of big-data science. In this scenario, a good candidate to substitute traditional polls are social media (Twitter in our study) which solve in one shot both the law response rate (million of people express their political preferences in the microblogging platform) and the social desirability biases. This is because social media users do not answer to any question, but freely express their ideas

in a social medium platform. However, one may argue that Twitter is generally bias towards young people thus providing a biased sample. Thus, a proper re-weighting of the data is needed, although the effects of re-weighting are expected to be less pronounced than in the polls of Elypsis. We introduce an AI model that builds up on previous work in [141] combining machine learning, network theory and big-data analytic techniques, that is able to overcome the problems presented so far and that correctly predicted the outcome of the 2019 Argentina primary and general elections.

4.3 Networks & Machine learning: AI based approach

The algorithm we propose improves upon previous work from [141] and consists of four phases (see Figure 22): data collection, text and user processing, tweets classification with machine learning and opinion modeling. While the first two phases are of standard practice in the literature, tweets classification by means of ML models only recently took place [141, 28], given the impossibility to classify by hand millions and millions of data. Opinion modeling, the core of our election prediction model, is an attempt to instantly capture people's opinion through time by means of a social network. To improve upon [141], we consider the cumulative opinion of people and define five prediction models based on different assumptions on the loyalty classes of users to candidates, homophily measures and re-weighting scenarios of the raw data. Below we explain each phase, highlighting the steps that make our full-fledge AI predictor a good candidate substitute for the traditional pollster methods.

4.3.1 Data collection

By means of the Twitter public APIs, we collected tweets from March 1, 2019 until October 27, 2019, filtered according to the following queries (corresponding to the candidates' name and handlers of the 2019 Argentina primary election): *Alberto AND Fernandez*, *alferdez*, *CFK*, *CFKAr-*

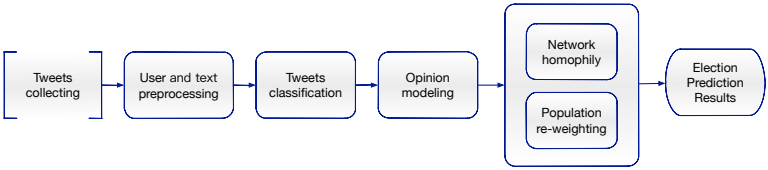


Figure 22: The flow of election prediction algorithm.

gentina, *Kirchner*, *mauriciomacri*, *Macri*, *Pichetto*, *MiguelPichetto*, *Lavagna*. Only tweets in Spanish were selected. Figure 23a shows the daily volume of tweets collected (brown line) while Figure 23b shows the daily number of users (green line). In blue we report the daily number of tweets/users which are classified, i.e. they posted at least one classified tweet. Users are classified with machine learning as supporters of Macri (Figure 23d, red line) if the majority of their daily tweets are classified in favor of Macri (Figure 23c, red line) or as supporters of Fernández in the other way around (blue line in Figure 23d and c). Hereafter we use FF to indicates the Fernández-Fernández formula and with MP we refers to the Macri-Pichetto formula (the outgoing president/vice-president candidate).

The activity of tweets/users shows a peak on August 11, 2020, i.e. the day of the primary election. In the period from March to October, we collected a daily average of 282,811 tweets posted by a daily average of 84,062 unique users. We daily classified 75% of these tweets and $\sim 76\%$ of the users (see Table B1 and Table C1 in the Appendix C). In total, by the end of October we collected around 110 million tweets broadcasted by 6.3 million users. This large amount of tweets collected has no precedent and is relevant in the light of considering that Argentina is one of the most tweeting per capita countries in the world.

4.3.2 User and text processing.

As a standard practice, raw data need to be elaborated before to be used. Here after we explain this step.

Bots detection. The identification of software that automatically injects

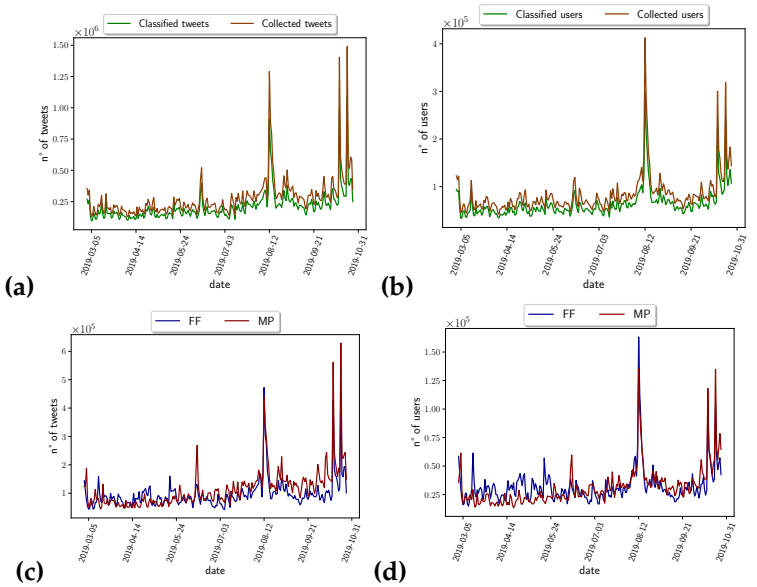


Figure 23: **(a)** Daily volume of collected (brown line) and classified (green line) tweets. **(b)** Daily volume of collected (brown line) and classified (green line) users. **(c)** Daily tweets supporting the FF/MP formula. **(d)** Daily users supporting the FF/MP formula.

information in the Twitter' system, is of fundamental importance to discern between " fake" and " genuine" users [167], the latter representing the real voters. According to [141] a good strategy is to extract the name of the Twitter client used to post each tweet from their source field and kept only tweets originating from an official Twitter client [28]. Figure 24a and b show the daily number of tweets posted by bots and the daily volume of bots, respectively. Figure 24c and d show the daily volume of classified tweets/bots. The daily average of bots between March and October is 732 with an overall daily activity (in average) of 2,243 tweets. The daily classified tweets are 1,617 while the daily classified bots are 560 bots. As for "genuine" users, a bot is classified if it share at least 1 classified tweet. In the entire dataset we found around 20,000 bots which posted 538,350 tweets. Let us notice that even though we classified the

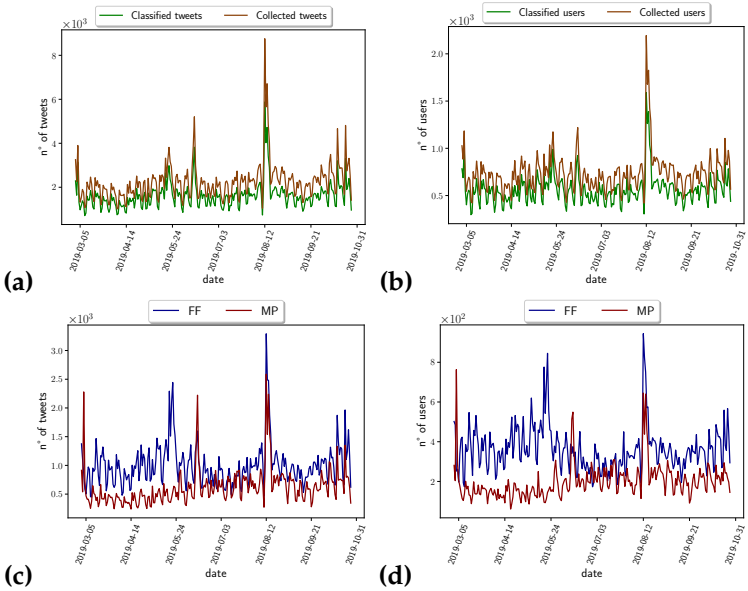


Figure 24: Bots analysis:**(a)** Daily volume of collected (brown line) and classified (green line) tweets. **(b)** Daily volume of collected (brown line) and classified (green line) users. **(c)** Daily tweets supporting the FF/MP formula. **(d)** Daily users supporting the FF/MP formula.

bots, they are not used for the final prediction since they do not corresponds to real voters.

Text standardization. Stop words removal and word tokenization are of common practice in Data mining and Natural language processing (NLP) techniques [168, 169]. For example, we keep the URLs as tokens since they usually point to resources determining the opinion of the tweet, through replacing all URLs by the token " URL ".

4.3.3 Tweets classification.

To build the training set we analyze the hashtags in Twitter. Users continuously labels their tweet with hashtags, which are acronyms able to directly transmit the user feeling/opinion toward a topic. We hand la-

beled the top hashtags used in the dataset (see Table C3 in the Appendix C). They are classified either as pro M(acri), F(ernández) or T(hird party) candidate, depending on who they support (with Third party we refer to the supporters of Lavagna, Espert and other secondary candidates).

Hashtag co-occurrence network. In order to check the quality of the classification of the classified hashtags we build the hashtag co-occurrence network $H(V, E)$ and statistically validate its edges [170, 141]. In the co-occurrence network the set of vertices $v \in V$ represents hashtags, and an edge e_{ij} is drawn between v_i and v_j if they appear together in a tweet. We test the statistical significance of each edge e_{ij} by computing the probability p_{ij} (p-value of the null hypothesis) to observe the corresponding number of co-occurrences by chance only knowing the number of occurrences c_i and c_j of the vertices v_i and v_j, and the total number of tweets N. Figure 25 shows the validated network. We only keep those edges with a p-value $p < 10^{-7}$. The blue community contains the hashtags in favor of Fernández, the red community those in favor of Macri and the green one (a very small group) are those in favor of the Third candidate. A look at the typologies of hashtags reveals the first differences in the supporters. Those in favor of Cristina Kirchner are much more passionate than the follower of Macri. For example, Kirchner's type of hashtags are #FuerzaCristina, #Nestorvuelva, #Nestorpudo or they are very negative to Macri as #NuncamasMacri. On the other hand, Macri's group is smaller and less passionate with hashtags like #Cambiemos or #MM2019 (see Figure 26), while support for the third candidate has not taken traction and its electoral base on Twitter is very small.

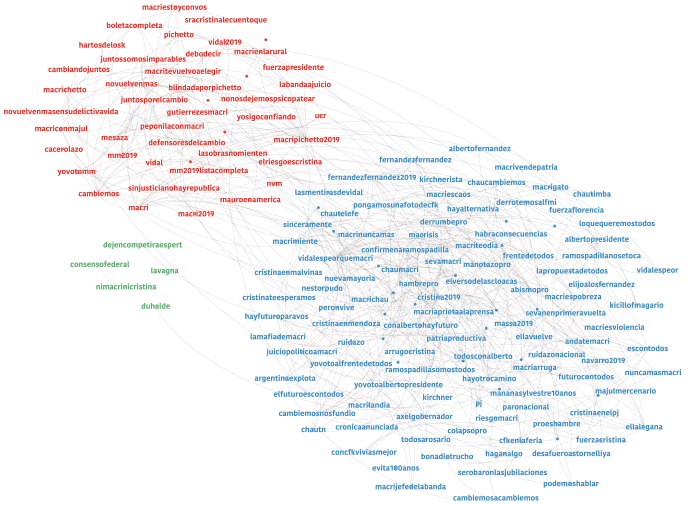
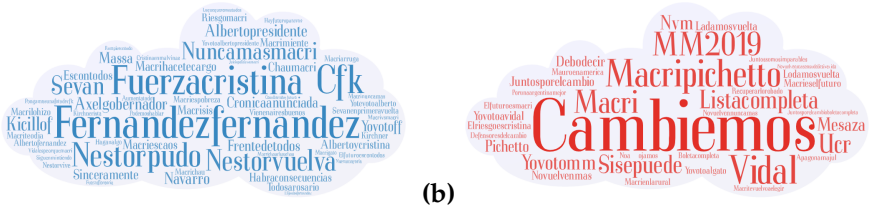


Figure 25: Hashtag co-occurrence network from March to August 2019. In blue the hashtags in favor of Alberto Fernández and Cristina Fernández de Kirchner, in red the hashtags in favor of Macri and in green those in favor of the Third party.



(a) **(b)**

Figure 26: Hashtag clouds. The dimension of the words is proportional to their frequency in the dataset.**(a)**The blue hashtags are the most frequently used in the tweets in favor of Fernández. **(b)** The red hashtags are those in favor of Macri.

In principle, counting the users and tweets according to the hashtags they use would predict the victory of Fernández over Macri. However this conclusion would be based only on $\sim$ 10,000 users (those expressing their opinion through hashtags). In order to get the opinion of all the users we train a machine learning model that classifies each tweet as AF, MM or Third party. (In what follows we also refer to the formulas FF for Fernández-Fernández and MP for Macri-Pichetto, the final formulas in the presidential contest). We use the previous set of hashtags expressing opinion to build a set of labeled tweets, which are used in turn to train a machine learning classifier. We use all the tweets (before August) which contain at least one of the classified hashtags to train the model. In the case of more than one hashtag for a tweet, we consider it only if all the hashtags are in favor of the same candidate. The use of hashtags that explicitly express an opinion in a tweet represents a "cost" in terms of self-exposition by Twitter users [171] and therefore allows one to select tweets that clearly state support or opposition to the candidates [141]. The training set consists of 228,133 tweets, i.e. the 0.33% of the total amount of collected tweets and the $\sim$90% of the hand-classified tweets (253,482 tweets). In order to find the best classifier we used five different classification models, the logistic regression (LR) with L_2 regularization, the support vector machine model (SVM), the Naive Bayes method (NB), the Random Forest (RF) and the Decision Tree (DT). All these models are validated on the remaining 10% of the classified tweets (25,349). Table 5 shows the results for the models. The logistic regression performs better

than the other models with an average group accuracy equals to 83%. Also recall and F1-score are equal to 83%. Support Vector Machine is the second classified, with an average accuracy of 81%. It follows the Naive Bayes with and average accuracy of 79.5%, the Random Forest and the Decision Tree.

Model	Precision (FF)	Recall (FF)	F1 (FF)	Precision (MP)	Recall (MP)	F1 (MP)
LR	0.83	0.83	0.83	0.83	0.83	0.83
SVM	0.81	0.81	0.81	0.81	0.80	0.81
NB	0.79	0.80	0.80	0.80	0.79	0.80
RF	0.74	0.80	0.77	0.79	0.72	0.75
DT	0.76	0.76	0.76	0.76	0.76	0.76

Table 5: Performance of the classification models: Logistic Regression (LR),Supporting Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) an Decision Tree (DT).

We recall that the logistic regression assigns to each tweet a probability p of belonging to a class. In our case such probability goes to one if the tweet supports Macri while it goes zero if it supports Fernández. As it is shown in Figure 27 the distribution of p contains two peaks, one on the left and one on the right, divided by a plateau. This is an encouraging result, since it proofs the efficacy of the model to discern between the two classes. We classify a tweet in favor of Macri if $p \geq 0.66$, in favor of Fernández if $p \leq 0.33$. Tweets with a value of p in the plateau are instead unclassified, meaning that the tweet does not contain sufficient information to be classified in either camp. According to this rule, in average we classify 211,229 genuine tweets and 1,617 "fake" tweets per day (see Table B1 and Table C1 in the Appendix C).

4.3.4 Opinion modeling

We can infer users' opinion from the majority of the tweets they post. Let $n_{t,F}$ be the number of tweets posted by a given user at time t in favor of Fernández and let $n_{t,M}$ be those supporting Macri. We define an instantaneous opinion over a window of length w and a cumulative average opinion as follow. In the first case, a user is classified as a supporter of Fernández (at a given day $t = d$) if $\sum_{t=d-w+1}^{d} n_{t,F} >$

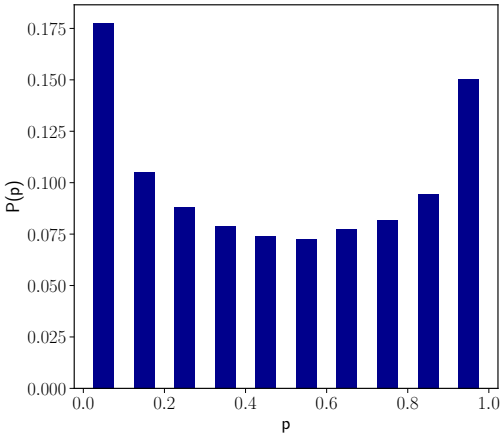


Figure 27: Probability distribution of p = probability of voting for Macri ($p = 0$ corresponds to Fernández) obtained by the Logistic Regression model.

$\sum_{t=d-w+1}^{d} n_{t,M}$, i.e if the majority of the tweets posted in the last w days were in favor of Fernández. The user is classified as a supporter of Macri if $\sum_{t=d-w+1}^{d} n_{t,F} < \sum_{t=d-w+1}^{d} n_{t,M}$. If none of the previous conditions is met, i.e. if $\sum_{t=d-w+1}^{d} n_{t,F} = \sum_{t=d-w+1}^{d} n_{t,M}$ then the user is classified as undecided. Let us notice that when w goes to one we have the 'most' instantaneous prediction, that is the prediction based on what people think in the last day. This instantaneous prediction model was used in Ref. [141] to match the results of the AI model to the aggregate of polls from the New York Times in the 2016 US election with excellent results. However, this predictor did not match the results of the electoral college, which required stratification by states. Thus, we further develop the AI model of [141] to add other predictors beyond the instantaneous measures. Traditional polls' data collection is an instantaneous prediction with a value of w that can go from few days up to few weeks, which is the time of collection of the poll data and this corresponds roughly to our instantaneous measurement above. However, the fact that we are able to track the same user over long period of time in Twitter allows us to extend the window of observation as far as we want to then define a new

measure that we call the cumulative opinion. The cumulative opinion in our model is defined by extending w to the initial date of collection for every time d of observation, i.e., $w = d$, thus considering the opinion of a user based on all the tweets he/she posted from time $t = 0$ upto the observation time d. That is, our prediction is longitudinal as we are able to follow the opinion of the same user over the entire period of observation of several months. In terms of traditional poll methods, a cumulative opinion would be obtained in a panel collecting for each respondent in the sample and for each day starting from $t = 0$ her/his preference toward a candidate. This possibility, which would require an unimaginable amount of effort and time for traditional poll methods, it is quite straightforward when it come to social media and big-data analyses.

We start by investigating the instantaneous response of the users in a fixed window of time. Figure 28a shows the Twitter supporters dynamics over time obtained with a window average, $w = 14$ days. Users are classified as MP (in red), FF (in blue) or Others (in green). Figure 28b shows the supporters dynamic (thick lines) compared with Elypsis prediction (thin dashed lines) without considering the undecided users in the normalization. In the same plots we also report the official results for both primaries and general elections. The comparison between the two pictures stands out as a approximate correlation between the Elypsis and the AI results for each candidate. However, in the comparison among candidates predictions may sometimes differs, as for example, right before the beginning of August, Elypsis gave as favorite MP while the AI instantaneous prediction was in favor of FF. Overall, as for the pollsters results, window average analyses are representative of the instantaneous sentiment of the people. As we see from the figures, instantaneous opinions are affected by considerable fluctuations [141] which make the prediction not reliable. In Figure C1 in the Appendix C we compare the average window opinion with other pollsters (Real Time Data, Management & Fit, Opinaia, Giacobbe and Elypsis). An interpolation (thin lines) shows similar trends as the AI-model window average, stressing that the conclusions made so far are more general then the simple comparison with Elypsis. In fact in [141] we have shown that the instantaneous

predictions of the AI model follows quite closely the aggregation of polls obtained from the New York Times, 'The Upshot', yet, it does not reproduce the results of the electoral college which requires a segmentation by states where proper prediction of rural and non-rural areas becomes the key and considering the cumulative opinion, not the instantaneous one, opinion of each users is crucial to correctly predict the elections.

Thus, we next study the opinion of each user by considering the cumulative number of tweets over the entire period of observation to define classify the voter's intention (MODEL 0). This cumulative approach takes into consideration the all the tweets together for each user since the first time they enter in the dataset and based the voter intention on all of them. This cumulative approach can only be done with Twitter and not with traditional polls, except for short times and particular cases as done by Elypsis before and after PASO.

Figure 29 shows the cumulative opinion from March 1 until a few days before the general elections. We can see that this approach captures the huge gap between the candidates, both for the primary election and the general election (vertical lines from the left to the right). While a low precision is of secondary importance when the difference between the opponents is high, it plays a central role when they have a close share of supporters. As an extreme example, in an almost perfect balanced situation the change of mind of just few people may flip the final outcome. If on the one hand a cumulative approach do reduce the fluctuations in the signal, it is also less sensitive to sudden change of opinion. A person can support a candidate until few days before the elections, for then change her/his mind because of some particular facts. This and other possibilities can be taken into account only by a model based on cumulative analyses, but able to capture the degree of loyalty of people towards the candidates over time. Differently from the traditional surveys, the real time data processing that underlies our AI algorithm gives the possibility to take into consideration this scenario. To understand how different re-weighting scenarios affect the results, below we introduce different loyalty classes of users towards the candidates and then we define several models matching the criteria previously discussed. These loyalty

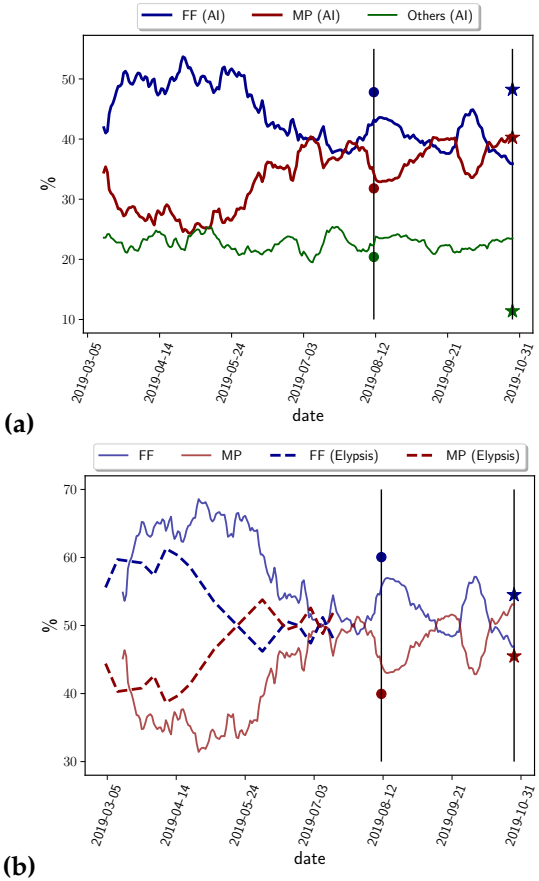


Figure 28: (a) Instantaneous prediction of the AI model obtained in a moving window of w=14 days. Vertical lines are (from the left to the right) the day of the primaries and general elections respectively. The circles represents the official results for the primaries while the stars those for the general elections. **(b)** Previous results compared with polls from Elypsis. Thick lines represent AI prediction while dashed line represent the Elypsis predictions.

classes can be only defined when we consider the cumulative opinion in a longitudinal study and cannot be investigated by traditional polls.

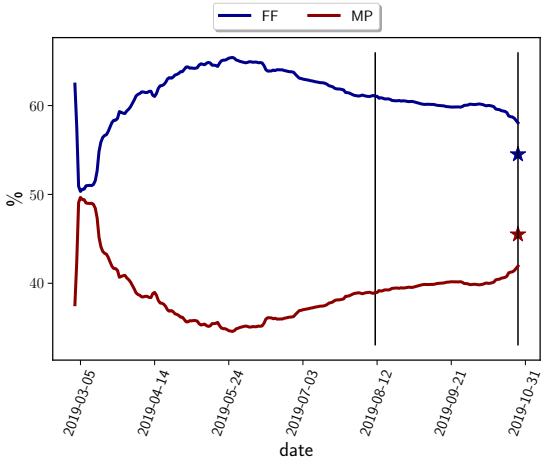


Figure 29: Simple cumulative predictions obtained without defining the loyalty classes (MODEL 0).

Loyalty classes. We define 5 classes of loyalty for users. Here we consider the MP supporters, but the definitions below similarly applied to the other candidates.

- Ultra Loyal (UL): users who always tweet only for the same candidate, namely $\sum_{t=T_0}^{T}(\frac{n_{M,t}}{n_{M,t}+n_{F,t}+n_{T,t}}) = 1$. where with $n_{x,t}$ we indicate the number of tweets that a given user post in favor of x, with $x \in \{$ Macri, Fernández, Third party$\}$.

Differently from the ultra loyal, which continuously post in favor of a candidate, the other classes take into consideration a possible change of opinion of a user. In order to detect sudden twist of opinions we focus on the classifications of the last k tweets posted by the users. We define:

- Loyal MP $\rightarrow$ MP: a user which is MP since the majority of tweet are for MP, but she/he also supported MP in the last k tweets. Mathematically speaking $\sum_{n=N-k}^{N} n_{M,n} > \sum_{n=N-k}^{N} n_{F,n} + n_{T,n}$. N is the total number of tweets posted by the user.

- Loyal MP $\rightarrow$ FF: users that are MP by the total cumulative count but they have tweeted for FF in the recent k tweets. In formula: $\sum_{n=N-k}^{N} n_{F,n} > \sum_{n=N-k}^{N} n_{M,n} + n_{T,n}$
- Loyal MP $\rightarrow$ TP: users supporting the third party in the last k tweets, i.e. $\sum_{n=N-k}^{N} n_{T,n} > \sum_{n=N-k}^{N} n_{M,n} + n_{F,n}$.
- Loyal MP $\rightarrow$ Undecided: all other individuals classified as MP but not included above.

Let us remind that unclassified refers to all those users who do not have any classified tweet. Figure 30 shows the cumulative prediction for each class, with T_0= March 1, 2019 and $k = 10$. The Ultra Loyal class for Fernández (FF) represents $\sim$ 33% of the populations while only $\sim$ 20% of the populations is Ultra Loyal towards Macri (MP). Loyal MP$\rightarrow$MP and loyal FF$\rightarrow$FF represents between the 8% and the 13% of the entire Twitter population. The percentage of the undecided is around 8% and the third party percentage. The other classes are close to 1 or 2%. In the next section we use these classes in order to define a better predictor.

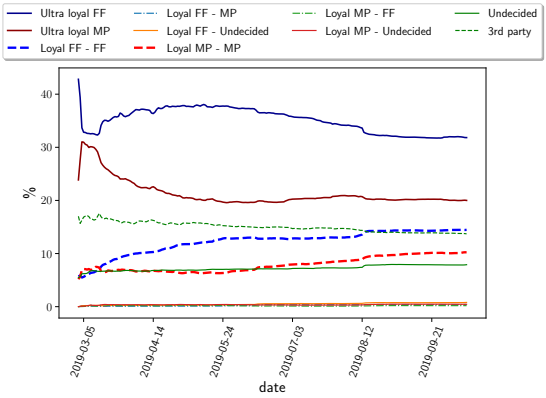


Figure 30: Cumulative users' opinion for each loyalty class over time.

AI Models. The loyalty classes introduced so far are one of the main differences with the other Twitter based studies: we use the machine learning classifier (logistic regression here) to define the loyalty of a user

and not to make predictions. We do that by grouping supporters as follows:

- Fernández supporters: all those users which are ultra loyal FF, loyal FF→FF, loyal FF→MP, loyal FF → Undecided.

- Macri supporters: all those users which are ultra loyal MP, loyal MP→MP, loyal MP→FF, loyal MP→Undecided.

In each group we put those users we are almost sure who they support because of their activity over time. However, as we saw in the previous section, undecided may play a central role in a scenario where few percentage points can flip the final result. Furthermore, understanding unclassified users (i.e. those users which do no not have any classified tweet) will also improve the final statistic. In order to take into account all the reasonable scenario we define three different models (starting from the classification in Fernández and Macri of above) and validate them against the final results of the election. Table 6 resumes the details of each model.

MODEL 1: Third party: Undecided→MP, Undecided→FF, Unclassified and Undecided→Undecided.

MODEL 2: Instead of simply grouping the undecided in a third party, we use network homophily to infer their political orientation. A user is classified as MP(Undecided) if the majority of her/his neighbors (in the indirected retweet network) support Macri. The same definitions applied for the other cases. In this model, FF(Undecided) are considered supporters of Fernández and MP(Undecided) supporters of Macri. Undecided(Undecided) and Unclassified belong to the third party.

MODEL 3: We collect the users' profiles, including the head portrait and the location information. Firstly users outsides Argentina are removed. using the face analyzing tool face++ [172], we obtain more than 400 thousand users' age and gender, see the population distribution of Twitter users from Figure C2. Following references [163, 164], we adjust the results of Model 2 by re-weighting the Twitter population to the Census data (see Appendix C).

Model	Supporters (Users)
MODEL 1	
FF	Ultra loyal MP, loyal MP→MP, loyal MP→FF and loyal MP→Undecided
MP	Ultra loyal FF, loyal FF→FF, loyal FF→MP, loyal FF → Undecided
Third party	Undecided→MP, Undecided→FF, Undecided→Undecided, Unclassified
MODEL 2	
FF	Ultra loyal FF, loyal FF→FF + loyal FF→MP + loyal FF→Undecided, FF(undecided)
MP	Ultra loyal FF, loyal FF→FF, loyal FF→MP, loyal FF → Undecided, MP(undecided)
Third party	Undecided(undecided), Unclassified
MODEL 3	
FF	FF as MODEL 2 + reweighting
MP	MP as MODEL 2 + reweighting
Third party	Third party as MODEL 2 + reweighting

Table 6: The definition of three models according opinion modeling.

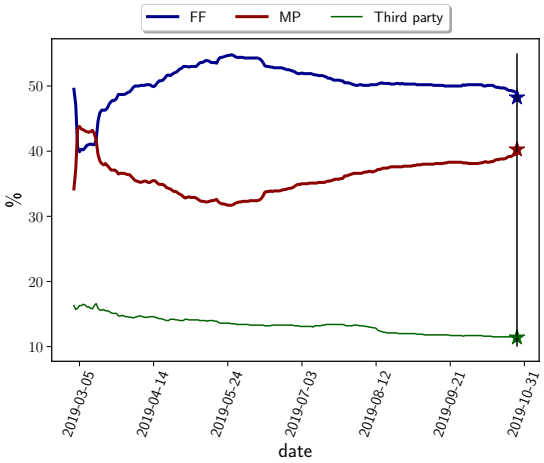


Figure 31: Cumulative MODEL 3 prediction of the AI model and comparison with primary on August 11, 2019 and general election results on October 27, 2019. Model 3 is the best fit to the real data.

4.4 Results

AI-based forecast: The Argentinian case. The models introduced so far allow us to define the daily supporters of each candidate according to their retweet activity. Indeed supporters are defined not simply according to the classification of the majority of their retweet, but on the basis

of the loyalty classes they belong. Similarly to the simple tweets classification, we can define for each model an instantaneous (window average) and a cumulative (average) opinion.

Figure 28 shows that an instantaneous indicator provides an approximate fitting to the results of polls. We have already used this indicator, in our previous study of the 2016 US presidential election, to precisely fit the New York Times Aggregator of Polls at The Upshot' [141, 173]. This aggregator unifies a thousands polls and weight them with proprietary information to produce a weighted average of all the most trustable pollster in USA. While this analysis is interesting and give the opportunity to predict instantaneous changes in electoral opinion, this indicator does not provide the electorate opinion as a whole and it is not the most important predictor of the election outcome. It is not the greatest information that can be extracted from social networks, either, and indeed, it failed to predict the US 2016 election and the present Argentina 2019. The estimator that predictor better the election is provided when we consider the cumulative number of users from the beginning of measurements, and not just the behavior of the users in a small window of observation.

For this reason here we directly focus on the cumulative prediction for the models introduced in the previous section. Table 7 reports the prediction of each model right before the day of the general election day: October 27, 2019. The official results saw the victory of Fernández with 48.24%. Macri scored 40.28% and the Third Party with 19.48%. The average predictions obtained by averaging the results of the five models are consistent (inside the standard error) with the official outcome. Indeed, we obtain (49.3 $\pm$ 2.1)% for FF, (36.8 $\pm$ 2.9)% for MP and (13.9 $\pm$ 4.8)% for the Third Party. This in an outstanding result which highlights the importance of considering loyalty classes for Political elections. In order to establish the best among the five models we compute the mean absolute error between each model' prediction and the final results. Let $Y = \{y_c\}$ with $c \in \{FF, MP, TP\}$ be the prediction of one model and let $X = \{x_c\}$ with $c \in \{FF, MP, TP\}$ be the official results. We define the MAE$_i$ (mean absolute error) for model i as $\frac{\sum_{i \in c}(|x_i - y_i|)}{3}$.

Table 7 shows the MAE for each model. MODEL 3, based on the ho-

MODEL	FF (%)	MP (%)	Third party (%)	MAE (%)
MODEL 1	45.9	32.5	21.6	6.71
MODEL 2	49.1	34.0	16.8	4.21
MODEL 3	48.9	39.6	11.5	0.53

Table 7: Models' prediction results for the general election on October 27.

mophily detection for the undecided is the best predictor with a mean absolute error of 0.53. This model predicted 48.9% for FF (an overestimation of 0.66 points if compared to the official result), 39.6% for MP (an underestimation of 0.68 points) and 11.6% for the Third Party. As a matter of fact, the AI model introduced so far is capable of predicting the Argentinian general elections, by giving a percentage of electors for each candidate close to the official one and outperforming traditional polls methods. See Figure 19b.

Maybe the most important result is the performance of our algorithm before the PASO, where all the pollsters failed too predict the + 16% points difference between the two candidates (by strongly underestimating their gap). MODEL 3 predicts a difference of almost 18% points in favor of FF, close to the official result.

While the PASO results appears of secondary importance, they play a central role in the Argentina political campaign and they are the most difficult to guess because it's the first time the citizens officially expressed their opinion on the election. Figure 19 shows how traditional pollsters modified their prediction after the primary elections, somehow fitting them with the PASO results. How they modified their predictions is still not clear and some of the pollsters (Elypsis for example) did not release any prediction after the PASO.

A study of the hashtags and queries of the followers of the FF formula indicates that the vast majority of the people focused more on the poor economic situation in which the country was instead of the judicial cases of corruption that affect the FF candidates. Most of the hashtags reflect sentiment of hunger, chaos, crisis and despair. On the other hand, the expression of the followers of Macri-Pichetto is reflected in hashtags

to give strength to the president but they do not reflect a feeling for the economic and political situation, but more a moral support, perhaps of resignation. The followers of Macri do not express too much their concerns about judicial cases of corruption either.

Finally, let us notice that the cumulative average depends on the initial time T_0. This value determines the initial fluctuations of the cumulative average, which generally stabilize into a value that it is difficult to change unless a big swing in opinion of the electorate. To investigate this effect, we have recalculated the cumulative average by changing the origin of measurement T_0 in Figure C3 in the Appendix C. As we see from this figure, the predictions for the general elections cluster around the same value. We use this fluctuations to compute the error associate to our final predictions. We define the error as the standard deviation over the results of different realizations with $t < T_0$. Regarding MODEL 3, the estimated average error is 0.53%. This result strengths the goodness of our prediction, consistent, inside the error bars, with the final results.

4.5 Discussion

One of the fundamental tools of artificial intelligence in social networks is that it captures changes in people's opinions without any intervention and for an extended time. Then AI can capture the sentiment of the millions of users who constantly express themselves on the internet and change or maintain their positions. AI can also filter this information from manipulators and bots and can reduce it to its essence, by overcoming the problems traditional pollsters face: low response rate, social desirability biases and the mis-representation of the population.

The results of our analyses show that AI applied to big-data can be used to successfully understand people' opinions over time. The possibility of following the opinion of the same people through time, and therefore the chance of defining loyalty classes is a fundamental step in order to make good predictions. AI allows both to get the percentage of supporters toward a candidate and reveals what is behind these numbers, giving an idea of people sentiments. This is of particular impor-

tance when one of the candidates is a controversial politician and can generate different feelings leading to strong polarization and biased responses to pollsters, which are not trusted anymore by the great majority of people.

We expect that in the future traditional surveys may be incrementally replaced by these new non-intrusive methods. AI is a thermometer that provides the key to predicting not only the elections but the great trends that develop at the local and global levels. We have shown how AI allows to synthesize the opinion of millions of people including those silent majorities of hidden voters who would not be heard otherwise. We must not ignore that people are tired of answering surveys. AI can then deduce, predict, interpret and understand what people want to express.

Chapter 5

Conclusion

In the first part of this work, we focused on one of Network Science's main findings. The scale-free power-law behavior of the vertex connectivity in many real-world networks. We argued that such behavior could be clouded by a sample limited in size. By employing finite-size scaling (FSS), we showed that we can not reject the scale-invariant hypothesis for many inherent structures of empirical networks. Similarly to the case of critical phenomena in statistical physics, we found a universal relation between the power-law exponent λ and the critical exponent d. Unlike statistical physics, where the power-law behavior results from critical phenomena, this may not be the case of scale-free networks. Critical phenomena assumes the presence of a parameter (in one dimension) that close to a critical value determines a jump from a predictable state (at equilibrium) to an unpredictable one, where the system is outside the equilibrium. The discussion of whether many empirical networks are at some critical point and how they would reach at that point is beyond the scope of this thesis. The effect of the critical exponent d on other properties of networks, such as the betweenness centrality or the cluster coefficient, is still unexplored. It may be possible to retrieve, again in analogy with the case of statistical physics, some scaling equations and a set of critical exponents. Together with the scaling equations these exponents would completely identify the system.

In the second part of the work, we focused on contact networks. We gave a probabilistic definition of contact. We implemented this definition to retrieve the chain of transmission of the COVID-19 in the metropolis of Fortaleza. The main finding is that the k-core structure works as a reservoir of the virus. Few people (the so-called super-spreaders) connect the reservoirs. Removing (quarantining) these spreaders would isolate the k-core structures and, therefore, the spread of the virus. We argued that this might be one reason why implemented strict quarantines around the world were not enough to bring the R_0 value below one. A micro-quarantine (quarantining the super-spreaders) should follow the general macro-quarantine. Our algorithm, which combines big data with advanced analyses taken from network theory, provides a possible strategy for the micro-quarantine to follow. We believe that the applicability is broader and not limited to quarantine strategies. For example, nowadays the main topic is vaccination. According to our results, super-spreaders are those who should be vaccinated first. This is indeed not far from common sense. For instance, according to our findings, doctors are those who should be vaccinated first. On the other hand, other targets are less trivial to define. Our algorithm would also help to identify other targets which should be prioritized.

Another line of research to follow would be to focus on the topological properties of the contact networks. Understanding the role of the critical exponents γ and d for different cities' contact networks can lead to fascinating results. For instance, the populations of cities could substitute the parameter N in the scaling function. Can the shape of the master-curve originating from different cities tell us something more than what we already know? Are contact networks of different cities self-similar? How demographic characteristics reflect on the structure of contact networks?

We conclude the thesis by considering another type of spreading, i.e., spreading of information on social networks. In particular, we use Twitter to infer the final results of the Argentinian political election of 2019. After investigating, we found that social desirability bias and low response rate are the main reasons for the failure of traditional polls

predicting the election results. These problems can be overcome using the enormous amount of data produced in social networks. Social media users freely express their opinion about a candidate or more general about a topic. By combining network theory and machine learning techniques, we build an algorithm that predicts the primary/general Argentinian election with high accuracy. The applicability of the proposed strategy is more general and can be applied to topics different from politics such as what people think about climate change or about a certain product.

In conclusion, we saw how technological developments led to new challenges, and these new challenges asked for new methodologies, tools, and theories. In the era of "big data," we found in network science a good ally. Applications of network science revealed a remarkable characteristic from social phenomena to economic ones. The features underlying many real-world phenomena are not entirely different from those characterizing physical phenomena such as phase transitions. As often happens in science, tools developed in the latter case can be adapted to the former ones, giving equal (or even better) performances. Obviously, this extends to all the disciplines. The fact that the common thread of many phenomena lies in their complexity is astonishing.

The amount of data we produce every day is giving us the possibility to understand this complexity better. All the studies we carry out are meant to understand ourselves (for example our brain or thoughts) and whatever surrounds us. Everything with the aim of making complexity less complex and our world more predictable.

Appendix A

Sampling bias, degree correlations and structure of the dataset

Here we report the steps to test the finite size scaling hypothesis of Eq. (2.2) together with the moments ratio test of Eq. (2.9). Note that in order to test Eqs. (2.3) and (2.11), one uses the number of edges E (e) associated with each (sub-)network of size N (n), and replaces d with d_E.

A.1 Finite Size Scaling analysis

Given an undirected network of size N, our analysis is based on the following steps.

1. We compute the degree distribution $p(k, N)$ and use the method of Clauset, Shalizi and Newman Clauset, Shalizi, and Newman [57] and Alstott, Bullmore, and Plenz [71] to estimate the best fitting power law parameters $\Gamma + 1$ and k_{min}.

2. We generate an ensemble of 100 sub-networks for each size $n \in \{\frac{N}{4}, \frac{N}{2}, \frac{3N}{4}\}$. Each sub-sample is obtained by picking n nodes at random from the original network and by deleting all the other

nodes and the links incident to them. We then compute the mean degree distribution $p(k, n)$ over each sub-network ensemble.

3. Both for the original network and for each sub-network, we check whether the (average) number of nodes n^* with $k \geq k_{min}$ is larger than $\ln N$. If this condition is not met, we classify the network as non scale-free and the analysis ends. Otherwise, we proceed by removing the region below k_{min} in both $p(k, N)$ and each $p(k, n)$, and renormalize them afterwards. As explained in the main text, this allows us to get rid of deviations at low degrees, including those induced by the sub-sampling (see also the Supplementary Information).

4. Using the moment ratio test, we determine d (and its associated error) as follows. We compute a given moment ratio $\langle k^i \rangle / \langle k^{i-1} \rangle$ on each (sub-)network of size n, and use least-squares to fit $\ln \langle k^i \rangle / \langle k^{i-1} \rangle$ versus $\ln n$. We then average the resulting fit slope over different choices of the moments (indexed by i) to obtain $-d$. Note that since this test is computationally less expensive than the collapse analysis (see below), we use more than four sub-network sizes. In particular we use 20 equally spaced values of $n \in [\frac{N}{4}, N]$, for each of which we compute the moments ratio (and associated error used as fit weight) over an ensemble of 100 n-sized sub-network built as described above.

5. For each (sub-)network size $n \in \{\frac{N}{4}, \frac{N}{2}, \frac{3N}{4}, N\}$ we obtain the cumulative degree distribution $P(k, n)$. We then determine the exponents γ and d (and their associated errors) that maximizes the quality of the collapse plot (see below). Notably, the scaling exponent d obtained from the collapse is always compatible with that obtained from the moment ratio test. Hence in order to decrease the computational cost of the method, one can in principle vary only γ while keeping d fixed at the value obtained from the moment ratios fit.

A.2 Dataset

We extract a collection of real network data from the Index of Complex Networks (ICON) at `https://icon.colorado.edu/` as well as the Koblenz Network Collection (KONECT) at `http://konect.uni-koblenz.de/`. The full list of networks we consider together with detailed results of the finite size scaling analysis are reported in the Supplementary Dataset Table. To define the dataset we select networks (removing duplicates appearing in both ICON and KONECT) according to the following criteria.

First, to allow for a reliable scaling analysis, we only use networks with $N > 1000$ and $E > 1000$ (for computational reasons, we did not consider networks with more than 50 million links). We then include undirected networks, as well as the undirected version of both directed and bipartite networks. Similarly, we consider binary networks as well as the binarized version of weighted and multi-edge networks. We however ignore networks that are marked as incomplete in the database. Importantly, among database entries that possibly represent the same real-world network we select only one (or at most a few) entry, and consistently we do the same for temporal networks (when there is only one snapshot, we ignore the time stamp of links).

In practice, in KONECT we select only the Wikipedia-related networks in English language. For ICON the implications are more profound. We ignore interactomes of the same species extracted from different experiments, the (almost 100) fungal growth networks, the (more than 100) Norwegian boards of directors graphs, the (more than 100) CAIDA snapshots denoting autonomous system relationships on the Internet, networks of software function for Callgraphs and digital circuits ITC99 and ISCAS89. We consider only one instance of Gnutella peer-to-peer file sharing network, as well as a few instances of the (more than 50) within-college Facebook social networks and of the (about 50) US States road networks. Among the (more than 100) KEGG metabolic networks, we select 17 species trying to balance the different taxonomies.

Thus, in our analysis, we do employ the same data source used by

Broido & Clauset Broido and Clauset [35], but we avoid over-represented network instances. As explained in the main text, this procedure removes the clustering of similar networks shown in Figure 7, and leads to less biased conclusions on the scale-free nature of networks belonging to different categories.

A.3 Degree cross-over k_c versus maximum degree k_{max}

The maximum degree in a network is defined as the value of k for which the probability to finding a node with equal or higher degree is $1/N$. For a power law degree distribution,

$$\int_{k_{max}}^{\infty} k^{-\lambda}\, dk \sim \frac{1}{N} \tag{A.1}$$

from which it follows that $k_{max} \sim N^{1/(\lambda-1)}$ Dorogovtsev, Mendes, and Samukhin [72]. The degree cross-over is instead the value of the degree for which the distribution has a crossover from a pure power law behavior to a faster decay (due to finite size effects). This is what defines k_c in the FSS ansatz. Using the same steps of Eq. (A.1) for $p(k, N) = k^{-\lambda} f(k/k_c)$ we get

$$\begin{aligned} \int_{k_{max}}^{\infty} k^{-\lambda} f(k/k_c) dk &= k_c^{1-\lambda} \int_{\frac{k_{max}}{k_c}}^{\infty} x^{-\lambda} f(x)\, dx \\ &= k_c^{1-\lambda} F(k_{max}/k_c) \sim \frac{1}{N} \end{aligned} \tag{A.2}$$

and again we find $k_c \sim N^{1/(\lambda-1)}$. Hence despite the values of k_{max} and k_c are different, both of them have the same scaling behavior with N. However in the main paper we showed that $k_c \sim N^{1/\lambda}$, at stake with Eq. (A.2). Moreover, Figure A1 shows that k_{max} also scales with $1/\lambda$ in our empirical data. Therefore k_{max} and k_c show the same dependence on N, but with an exponent different from the expected one. We believe this is due to the correlations inside the networks which alter the dependence on N Boguñá, Pastor-Satorras, and Vespignani [80].

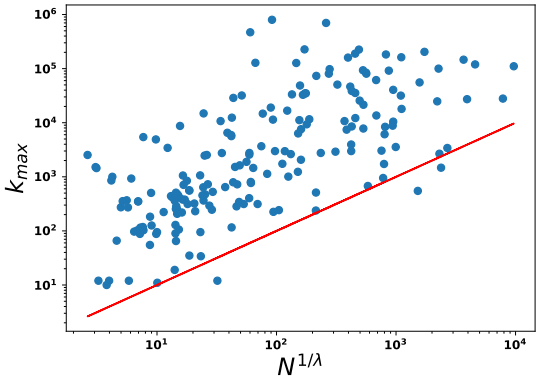


Figure A1: Empirical relation of k_{max} versus $N^{1/\lambda} = N^{1/(\Gamma+1)}$ for the real networks in our dataset. Here Γ is the exponent estimated via maximum-likelihood fitting Clauset, Shalizi, and Newman [57]. The red line represents the identity and serves as a guide for the eye.

Another argument in favor of the same scaling of k_c and k_{max} is as follows. By definition, k_c is the value of k for which the argument of the scaling function becomes constant, whence $k_c \sim N^{-d}$. Concerning k_{max}, since the cumulate of the distribution of maxima is

$$P^{>}_{max}(k, N) = Nk^{-\gamma} f(kN^d) \tag{A.3}$$

for not too small k, we have that the characteristic maximum, defined as $k^*_{max} \equiv \langle k^i \rangle_{max} / \langle k^{i-1} \rangle_{max}$, where $i > \gamma + 1$ and $\langle \cdot \rangle_{max}$ is the average with the distribution $-\frac{d}{dk} P^{>}_{max}(k, N)$, behaves as $k^*_{max} \sim N^{-d}$ for large N.

A.4 Finite-size scaling and system size

In order to apply finite-size scaling to network data we start from an empirical network of N nodes and then build smaller sub-networks through a sampling scheme (see below for more details). Given the values of N typical of empirical networks, the down-scaling of the system is performed linearly, whence the sub-networks of sizes $\frac{N}{4}$, $\frac{N}{2}$ and $\frac{3N}{4}$ considered in this study. However, when considering artificial network data we

are free to down-scale the system even by orders of magnitude. This is shown in Figure A2 in the case of a Barabási-Albert model with $N = 10^5$. Notably, parameters estimated on this systems are perfectly in line with those estimated with the linear sub-sampling scheme (see Chapter 2 and Figure A5).

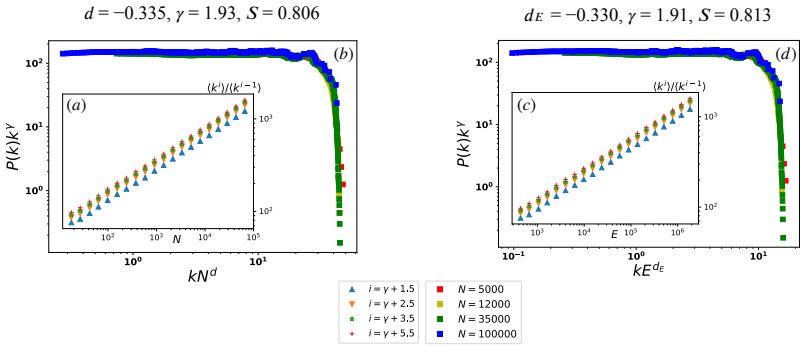


Figure A2: Scaling analysis on a numerical realization of the Barabási-Albert model with $N = 10^5$ and $k_{min} = 14$. Panels (a), (b): the scaling analysis with N yields $d = -0.335 \pm 0.024$, $\gamma = 1.93 \pm 0.03$ and $S = 0.81$. Panels (c), (d): the scaling analysis with E yields $d_E = -0.330 \pm 0.024$, $\gamma = 1.91 \pm 0.03$ and $S = 0.81$.

A.5 Node sampling method

The down-scaling of a network, namely how to derive a representative sample of the original network, is still an open problem nowadays. There are two main approaches in the literature to deal with this issue: sampling Leskovec and Faloutsos [174], Stumpf, Wiuf, and May [69], and Lee, Kim, and Jeong [70] and renormalization García-Pérez, Boguñá, and Serrano [98], Song, Havlin, and Makse [99], Gfeller and De Los Rios [175], and Serrano, Krioukov, and Boguñá [176]. Despite their simplicity, sampling procedures do not necessary preserve the properties of the original network. On the contrary, renormalization (i.e., coarse grain) procedures are based on the conservation of these properties during the down-

scaling; however these methods lack of generality by requiring ad hoc assumptions. In this paper we decide to use the most general, although not necessary the most accurate, scheme: random node sampling. Here we provide a detailed discussion on the degree properties of our sampled networks. To build a sub-network instance, a set of n nodes (with $n \leq N$) are selected uniformly at random. The sampled network is the network induced by these nodes and by all the links among them. See Figure A3 for an sample illustration. Figure A4 reports the degree distributions of the (sub-)network in the case of a Barabási-Albert model with $N = 10^4$ and various densities ($\langle k \rangle = 1, 5, 14$ for the first, second and third column respectively). As in the main paper, we chose $n \in \{\frac{N}{4}, \frac{N}{2}, \frac{3N}{4}, N\}$ as possible size of the sub-network. However, different sizes of the network may be chosen. Indeed, as we show in Figure A2, results do not change by changing the sub-networks dimension. In the first row of the figure (panels $a - c$) we report the cumulative degree distributions for a single instance of a sub-sampled network. In agreement with Stumpf, Wiuf, and May [69], for nodes with low connectivity and for $n \ll N$ the random node sampling does not preserve the shape of the power law degree distribution. However these deviations are reduced by increasing the average degree of the network (i.e., by moving from the left to the right panel). In the second row of the figure (panels $d - f$) we report the cumulative degree distributions obtained by averaging over 100 instances of the sub-sampling. As expected, fluctuations in the degree are drastically reduced. Notably, in the region between k_{min} and k_{max} where the scaling assumption holds, the original degree distribution is preserved. This is particularly evident when $\langle k \rangle > 1$. However, because the down-scaling changes the mean degree of the networks, these curves do not collapse on top of each other as one would expect. This issue also affects renormalization procedures Serrano, Krioukov, and Boguñá [176] and García-Pérez, Boguñá, and Serrano [98], because one should compare *rescaled* quantities. To do that there are two options. The first is to work with rescaled degrees $k/\langle k \rangle$ instead of actual degrees Serrano, Krioukov, and Boguñá [176]. As shown in the third row of the figure (panels $g - i$), by doing this all the distributions follow a single master curve for

$k \geq k_{min}$. The other possibility is to adjust the average degree without modifying the statistical properties of the network García-Pérez, Boguñá, and Serrano [98], for instance by taking the part of the distribution beyond k_{min} (estimated by maximum-likelihood power law fitting on the whole network Clauset, Shalizi, and Newman [57] and Alstott, Bullmore, and Plenz [71]) and renormalizing it such that $\sum_{k=k_{min}}^{k_{max}} P(k) = 1$. By doing this, as shown in the fourth row of the figure (panels $j - l$), the distributions for different n still collapse on the top of each other (without the need to rescale by the mean degree), apart for the region near k_{max} where finite size effects step in. Notably, the mean degree does not change as an effect of the cut of the distribution below k_{min}. Indeed, let us denote by n^* the number of nodes with $k \geq k_{min}$, and by e^* the number of links connecting these n^* nodes among themselves (l links) plus the number of links connecting these n^* nodes with the other nodes in the network with $k < k_{min}$ (s stubs). For scale-free network with $\lambda > 2$, the mean degree is finite and since $\langle k \rangle = 2E/N \sim$ constant we have $E \sim N$. We measured n^* versus e^* for 20 equally spaced values of n. A fitting of $\ln e^*$ vs $\ln n^*$ yields a slope equal to 1.02 ± 0.10, in agreement with the expected value of one obtained when the mean degree is preserved during the down-scaling of the network.

In conclusion, as long as the considered network has an average degree much bigger than one (as in our case, see the left panel of Figure A5, the sub-sampling procedure preserves pretty well the degree distribution of the mother network. However, we remark that our random sampling procedure cannot substitute more rigorous renormalization methodologies, since it does not allow us to have a unique representation of the reduced network. Finally let us notice that the reason why we use the cut procedure is because we are interested in computing the quality of the collapse only in the region where the scaling hypotheses holds, i.e., when $k \geq k_{min}$.

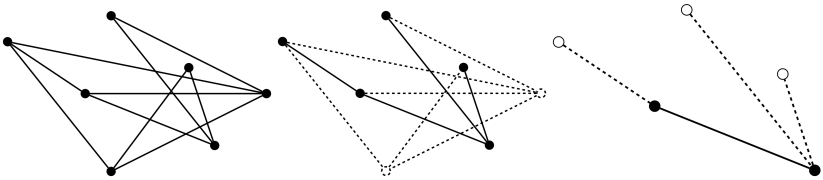


Figure A3: Illustration of the sub-sampling scheme. The original network is reported on the left. In the middle we report the sub-network induced by 5 random selected nodes, with removed links/nodes represented with dashed lines. On the right an example of the counting procedure after the cut at k_{min}. If $k_{min} = 2$, then $N^* = N_{k \geq k_{min}} = 2$ (filled circles) and $E^* = l + s = 1 + 3 = 4$, with l number of links (solid lines) and s number of stubs (dashed line).

A.6 Choice of the average degree

In the main paper we study the benchmark Barabási-Albert model with $m \equiv \langle k \rangle = 14$. We choose this value in accordance with the mean of the $\langle k \rangle$ in the empirical network of our dataset (see panel a of Figure A5). Other choices of m are however possible. Panels b and c of Figure A5 show the scaling analysis with N and E, respectively, for a Barabási-Albert model with $m = 5$. The power law exponent estimated by means of the KS test is $\Gamma = 1.847 \pm 0.039$. In the first case the quality of the collapse is $S = 0.26$. $\gamma = 1.807 \pm 0.108$ while $d = -0.378 \pm 0.133$. In the second case we have $S = 0.25$, $\gamma = 1.807 \pm 0.113$ and $d_E = -0.361 \pm 0.119$. The network is classified as strongly scale-free. Note that the quality of the collapse S does not depend on the average degree of the network (both values $m = 5$ and $m = 14$ give a close result for S). Moreover, the value of the estimated exponent γ is different from 2 in both cases. This is not a consequence of using a large m, because $\gamma = 2$ holds in the asymptotic limit $N \to \infty$ while the number of nodes in our networks is always finite.

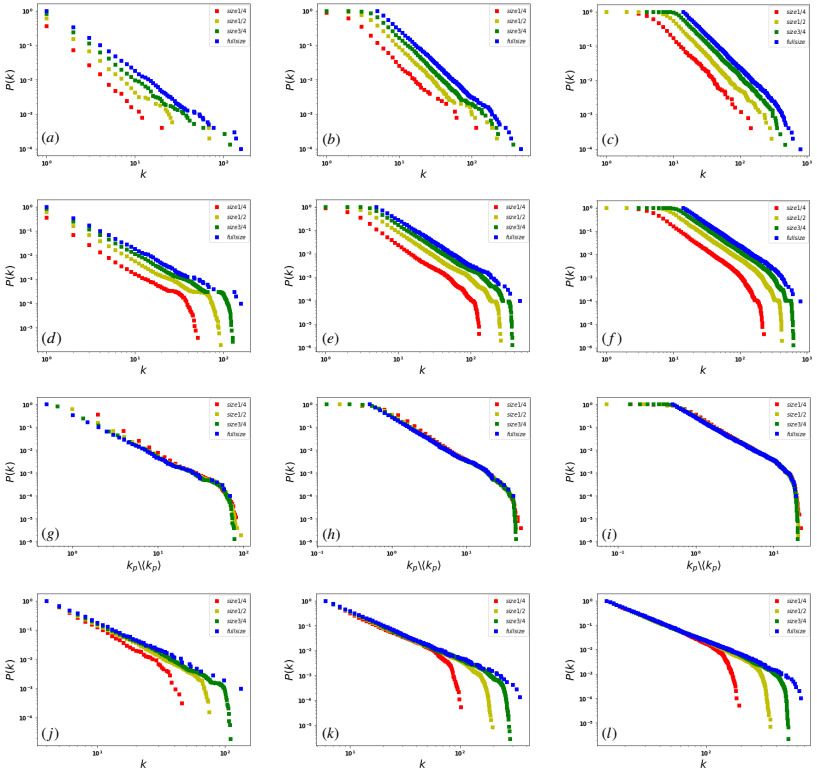


Figure A4: Analysis on three different realizations of a Barabási-Albert model with $N = 10^4$ and $\langle k \rangle = 1, 5, 14$, from the left to the right. On the top the degree distributions extracted using single sub-sample. In the second row the results for multiple sampling for each value of p. The third row reports the same results of the second row, but with rescaled degree sequences. On the bottom the result with multiple sampling for each value of p after the cut and the normalization.

A.7 Clustering in the γ-S plane

Panel a of Figure A6 shows the scatter plot of γ versus S, with each point representing an empirical network. As mentioned in the main text, we observe no clusterization of networks amenable to categories. This hap-

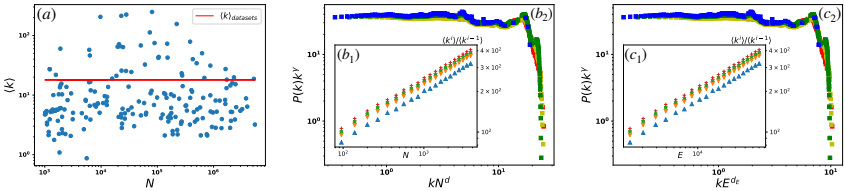


Figure A5: (a): Mean degree $\langle k \rangle$ as a function of the number of nodes N. Each point corresponds to an empirical network in our dataset, while the red line is the value $\langle k \rangle = 14$ we use in the benchmark Barabási-Albert model. The other two panels (b, c): report the scaling analysis with N and E in the case of a Barabási-Albert network with $m \equiv \langle k \rangle = 5$.

pens however because we built the dataset balancing the various network categories. Networks that are very similar may instead cluster together, as shown in panel b of the same figure where we report all the 109 KEGG protein interaction networks that can be found on ICON.

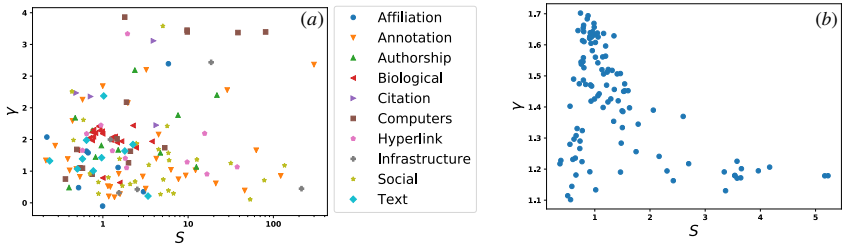


Figure A6: Relation between the power law exponent γ and the quality of the collapse S resulting from the finite size scaling analysis with N. Each point represents an empirical network. Panel a reports results for the whole dataset at our disposal (all the networks for which S is defined), whereas, panel b focuses only on metabolic networks of various species from KEGG [177] – which are mostly not included in panel a.

A.8 Binning errors

As explained in the main paper, in order compute the quality of the collapse S we need to assign an error $\sigma(k)$ to each binned value of $p(k)$ in

each (sub-)network. When degrees are independent and not correlated, then the counts in each bin k should closely follow a Poisson distribution. In this case a good estimate for the error in each bin is simply the square root of the counts: if $p(k) = n_k/n$ is the fraction of nodes with degree k in a set of n nodes, then the error is $\sigma(k) = \sqrt{n_k}/n$. However in real networks degrees are typically correlated, therefore this approach could lead to uncontrolled biases. An alternative way that accounts for correlations is to compute the errors by means of a bootstrapping approach. Given a degree sequence of length n we uniformly at random extract $\tilde{n} \leq n$ values (extractions with replacement). By repeating this procedure an arbitrary number of times and by computing the standard deviation of the counts in each bin among all the realizations we obtain an unbiased estimation of the error.

Panel a of Figure A7 shows the scatter plot of the errors obtained using the two described approaches, in the case of a numerical realizations of the Barabási-Albert model with $m = 14$ and $N = 10^4$. In particular we consider the sub-network of size $3N/4$, whose degree distribution is obtained by averaging over multiple (100) sub-samples. Hence for the bootstrap case we have a degree sequence of total length $n = 100 \times 3N/4$, from which we generate 10^4 bootstrapped sequences of length $\tilde{n}$. We use $\tilde{n} = n$ (remember that we allow for multiple extractions of the same entry). The two errors are very similar (linear fitting returns a slope close to one). Hence the two approaches are almost equivalent – for computational reasons we used the Poissonian approach. Panel b of the figure further shows that the higher difference between the two errors is obtained especially for high values of k. At last in panel c we show that the average ratio of the Poissonian and bootstrapped errors is inversely proportional to $\tilde{n}$, but stays in the range $[1, 2]$ until very low values of $\tilde{n} \sim n/10$.

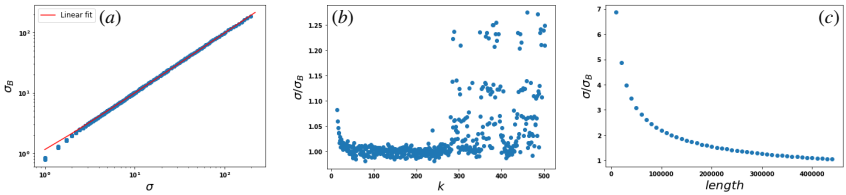


Figure A7: (a) Scatter plot of σ_B (the error computed with the bootstrap procedure) versus σ (the Poissonian error obtained as the square root of the counts), in the case of a Barabási-Albert model with $m = 14$ and $N = 10^4$. We use $\tilde{n} = n$. The linear fit (shown as a red line) has a slope equal to 0.966 ± 0.001. (b) Ratio σ/σ_B as a function of the binned degree k for $\tilde{n} = n$. (c) Mean ratio σ/σ_B as a function of the length $\tilde{n}$ of the bootstrapped sequence.

Appendix B

Sampling bias, robustness and incompleteness of the dataset

This section addresses concerns regarding the extent to which the present results would hold under implementation in terms of robustness to data quality and coverage, sampling bias on demographics such as coverage of location, socio-economic status, age and gender and privacy. Some of these issues are discussed elsewhere [178, 122].

B.1 Datasets

The GPS dataset from "Grandata-United Nations Development Program me partnership to combat COVID-19 with data" [46] provides the geolocation data of users of a compilation of hundreds of mobile apps across Latin America. Each datapoint registers the geohash location of a mobile ID with a 12 digits precision (cm resolution, although we use meter resolution due to noise in the determination of the exact GPS location). The mobile ID is MD5 hashed data associated with a hash of the MADID (Mobile Advertising ID).

See [46]: The full GPS dataset covers 10 countries across Latin Amer-

Month 2020	Unique users	Daily users	GPS Datapoints
March	111168	43231	153282989
April	145148	49913	161304623
May	158405	63489	221381875
Total	257606	52211	535969487

Table B1: GPS datasets of Ceará from [46]

ica: Argentina, Brazil, Chile, Colombia, Ecuador, Guatemala, Mexico, Paraguay, Peru and Uruguay. Average number of users in the dataset per day: Uruguay: tens of thousands of unique users per day. Argentina: hundreds of thousands of unique users per day. Mexico: Millions of unique users per day. Brazil: tens of millions of unique users per day. Over the month of March, lockdowns have been imposed on each country. Pre-quarantine dates in Latin America range: March: 1st through 19th, 2020. Quarantine in Latin America dates range: March 20th till today, as of June 9.

We measure the average root mean square displacement (MSD) of the users in each country per day. Averaging over all the countries in Latin America we measure a -53.4% reduction in the MSD from a reference date before lockdown on March 6, 2020 to May 1, 2020, see [46].

In the state of Ceará, we find a reduction in mobility from MSD $=$ 988m pre-quarantine to MSD $=$ 430m, giving a reduction of 56.5%. The GCC of transmission contacts defined over one week window in the state of Ceará is reduced from a size of 9402 users pre-lockdown to an average size of 983 users per week during the lockdown, giving a reduction of the GCC to -10.4% from its original pre-quarantine size.

Patient's dataset is provided by the Department of Epidemiological Surveillance of the Fortaleza Health Secretariat in the city of Fortaleza, Ceará, Brazil. The research was approved by IRB at CCNY and UFC. Patient data was used with the approval and consent from the Epidemiological Surveillance Department, Fortaleza Health Secretariat at the Prefei tura de Fortaleza, Ceará, Brazil.

We cross the information from the Health Department dataset with

the GPS dataset detecting the mobile ID's that are related to any confirmed case from the Health Department data using their geolocalized address. We set the night time period from 10 PM to 5 AM as a time window with high probability of being at the location. Afterwards, we identify the mobile ID of the user that spent more than 15% of the total time in the location in those time-spatial areas during March-May 2020. We assign that mobile ID to the patient.

In the Fortaleza dataset there were 8323 cumulative cases from March 1 to May 7, 2020, out of which 5814 contain geolocation and date of first symptoms, and 1440 cases were matched to the GPS dataset. Lockdown started on March 19, 2020. On the month of March, 465 cases are matched which are used to calibrate the model. The number of nodes in the transmission tree of Figure 9c is 90, the number of connections in the tree is 60, and the number of secondary infections in the tree is 52. Our results can be applied to other areas as well. In preliminary studies we find that they are reproducible in other studied regions such as the state of Puebla, México.

B.2 Model calibration

In order to calibrate the contact model, we look for the combination of parameters r, T and p_c that better match the basic reproduction number $R_0 = 2.78$ obtained by fitting the cumulative number of cases with an SEIR model. The calibration period runs over the month of March, 2020 in the state of Ceará. In March 2020, there were 1392 infected cases reported localized in the city of Fortaleza. Out of these cases, we cross-checked 465 infected users in the GPS dataset. We then trace the contacts among this 465 infected users over the infectiousness period of -4d/+7d from date of first symptoms.

We run the calibration over a set of hyperparameters to search for the best set that most closely replicate the basic reproduction number $R_0 = 2.78$. The closest fit is obtained for: T= 30 min, r =8 m, and $p_c = 0.9$. Using these hyperparameters, we find that 90 unique infected users participated in contact events with other infected users. That is,

90 users have either a non-zero in-degree or out-degree or both. The remaining 375 users had neither in- nor out-degree detected in the GPS database. The 90 users are plotted in the transmission tree in Figure 9c. The tree contains 60 contact events, of which only 52 unique are unique contacts. This result leads to an average out-degree for the set of located infected users of $\langle k_{\text{out}} \rangle = 52/465 = 0.112$. Despite the low value of $\langle k_{\text{out}} \rangle$, we notice that our result is based on a subsample (111168 users) of the entire population of Fortaleza (2643247 citizens). Therefore $\langle k_{\text{out}} \rangle \propto CxR_0$, with $C = 111168/2643247 \sim 23.77$. This is indeed consistent with our finding.

The distribution of probabilities $P_i[n]$ shown in Figure 10a appears to be extremely polarized leaving the p_c with a wide range of optimal values for an appropriate fitting. We use p_c = 0.9 which fits our results and at the same time is large enough to filter irrelevant noisy contacts (the results do not depend largely on p_c). As shown in Figure 10b, $T = 30$ min is also the value at which the average of the contact probability $\langle P_i[n] \rangle_T$ starts to decrease with T. Therefore, $T = 30$ min is the smallest value of a well-behaved T, since we expect that the contact probability should decrease with T. Eight meters is also consistent with typical precision of geolocation given by noise level.

Sparseness of the dataset Due to the sparseness of the GPS data, we do not have access to all the contact between the infected people. To account for the smaller coverage of the GPS data in the calculation of R_0, we first obtain an effective $R_0' = 0.112$ valid for the smaller GPS dataset. This number obtained for the GPS dataset should be rescaled by the population ratio between the real population and the GPS sample, which is a factor of 23.77 and provides $R_0 = 2.66$, which is consistent with the values directly measured from the data.

It is important to note that this rescaling is only used to estimate the epidemiological parameter R_0 and does not in principle affect the further modeling of the contact networks, except for the fact that we use this estimated value R_0' to estimate the hyperparameters of the model (T, r). This means that, using other GPS datasets with different coverage, a new set of parameters needs to be determined to fine tune the model to the

particular coverage of the GPS dataset used. Therefore, in a less sparser dataset in an actual application with larger number of mobile app users, the parameters of the model (T, r, p_c) should be calibrated accordingly.

We note that in our rescaling, we do not rescale the network, nor the in- or out-degree distributions, but just use the smaller GPS sample to obtain the hyperparameters of the model.

B.3 Network reconstruction

The infected people have been informed to quarantine since they are infected, tested and with symptoms and have been informed by the medical authorities to isolate. Furthermore, since our data have been acquired under the most strict mass quarantine imposed by the authorities, everyone in the city of Fortaleza has been already informed to quarantine, including the contacts of the infected people. Indeed, the GCC is reduced to 10% and the mobility to 40%. Despite all these isolation efforts the pandemic persists. Our comprehensive contact tracing algorithm identifies further k-core superspreader events which kept the spreading.

The crucial feature of COVID-19 is that the peak of infectiousness occurs before the onset of symptoms, as shown in Figure 9a [117]. Thus, after a person reports symptoms or tests positive, contact tracing goes back in time to capture past contacts in the past GCC.

For instance, by capturing the contacts today and in the past, since there is a delay of up to 12 days of incubation and between 3 to 10 days to infectiousness (latency, see Figure 9a), then we are effectively stopping the transmission chain "in the future". This situation is further clarified in Figure B1. The infected person produces an exposure at first contact today or in the past, labeled as "Potential Target 1st layer" in the figure. This first layer contact then enters into the infectious period anywhere between 3 to 10 days later, and produces a second layer contact as shown in the figure, thus "propagating" the contact network to the future. These contacts "in the future of today" are considered in the network only when we move forward in time the window of observation. That is, they are considered when the "future contacts" are already in the

past of the observation time.

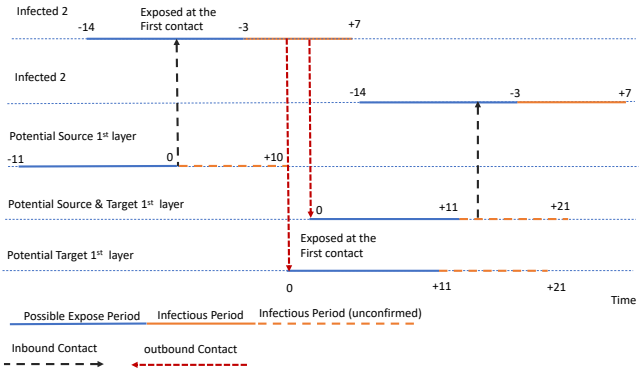


Figure B1: Contact layers of pre-symptomatic and asymptomatic captured by the model. Our treatment of asymptomatic cases is to increase the exposure period to -14 days to accounting for possible two-chains of infection as shown in the figure. Contacts between -2 days to -14 days from the day of first symptoms are more likely to be an exposure from an asymptomatic infected person. Contact from -2 days to +7 days from first symptoms are considered to be transmissions contacts from the patient.

When the infected person reports the symptoms, even if she/he enters into strict isolation, he/she already infected most of her/his contacts. The same happens for the contacts of the contacts and so on. The situation is worst due to the asymptomatic cases (which are captured by our model). People who get sick in the future have been infected in the past and one needs to reconstruct the network of contacts using these information of past contacts which then propagates into the future as time moves on.

The optimality in our approach is obtained by comparing the different strategies to dismantle the GCC. While other more optimal interventions could be devised by taking into account the fact that intervening on nodes active initially is better than intervening on nodes more active at a much later point, strategies following the contacts and applying interventions as the contacts are created may be difficult to implement in real

time. Therefore, our approach is not the optimal quarantine, but represents an optimized strategy to dismantle the GCC as compared with the rest of the network centralities studied. In fact, applying interventions one by one, may lead to even more efficient quarantines than those found here.

B.4 Sampling bias

Uniformity of data coverage of the dataset across socio-economical classes, age groups, and geographical regions may affect our results due to the relatively low coverage of the GPS data. This problem is important given that the sample of the population is relatively small compared to the underlying population: 111168 users out of the total population in the city of Fortaleza of 2643247 at 4% of the population and the biases could be substantial. Further, this problem is important for COVID-19 which itself has non-uniform incidence across these factors. Thus, we have investigated the sampling bias of our GPS population `https://en.wikipedia.org/wiki/Sampling_bias`.

We investigate the most likely bias, i.e. geographical coverage, socio-economic status (wealthier over-represented) and age and gender of users (younger over-represented) since one would expect the majority of symptomatic cases are in the lower end of the socio-economic spectrum, and older age groups. We quantify these biases in the GPS dataset from apps by assigning each mobile app user to a geolocalized residential area defined as the place where the user spends most of the time at night between the hours of 10 PM and 5 AM in the period of study. Using these data we study the distribution of geographical localization of the app users. We consider the 120 neighborhoods (quarter or 'bairro' in Portuguese) defined by the administrative boundaries in Fortaleza. Since the neighborhoods are extensive, the geolocalization of the users is non-identifiable. The population of each neighbourhood is provided by the Instituto Brasileiro de Geografia e Estatistica (IBGE or Census Bureau). By using the geolocation of each GPS user we calculate the fraction of app users in each neighborhood and then compare with the real fraction

of the population of each neighborhood obtained from IBGE. The distribution of the populations obtained from GPS data and the real population distribution from IBGE are shown in Figure B2. We perform a two-sample Kolmogorov-Smirnov (KS) test and find p-value =0.388, KS distance = 0.117, indicating that we cannot reject the hypothesis that the GPS data and the real data come from the same distribution. Therefore, we conclude that the GPS data has an acceptable geographical coverage of the real population indicating no sampling bias in geolocation under a statistical test.

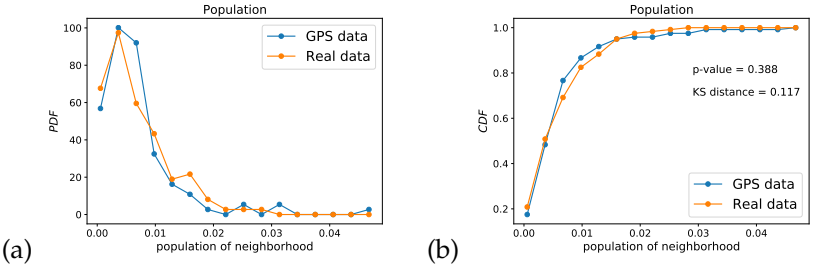


Figure B2: (a) Probability density function and **(b)** Cumulative distribution function of the fraction of the population per neighborhood in Fortaleza to the total population. We show the real distributions and the distributions from the apps GPS data. Both distributions pass a two-sample KS test indicating that we cannot reject the hypothesis that they come from the same distribution under the test.

Furthermore each neighborhood has a distinct Human Development Index ($0 \leq$HDI≤ 1) provided as well by IBGE. By using this metric of socio-economic status, we study the possibility of socio-economic bias in the population sample. We now cluster the neighborhoods by their HDI and plot the PDF and CDF of the population of neighborhoods (measured as the fraction to the total population) with a given HDI obtained from the GPS data sample and from the real population from IBGE. Results are shown in Figure B3. We performed a two-sample Kolmogorov-Smirnov test and find that the sample distribution of HDI obtained from the GPS data and the real population distribution of HDI socio-economic status pass the KS test, p-value $= 0.699$, KS distance = 0.250. Thus, we

cannot reject the hypothesis that the GPS and real data come from the same distribution indicating lack of sampling bias

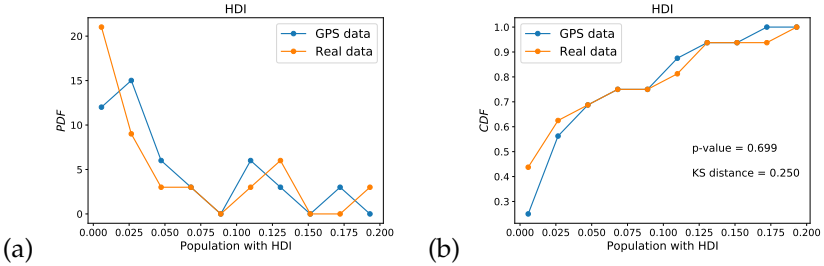
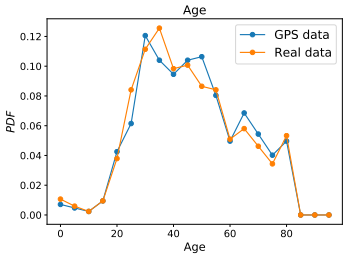


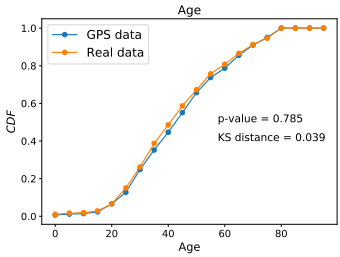
Figure B3: (a) Probability density function and **(b)** Cumulative distribution function of the fraction of the population per neighborhood with a given HDI in Fortaleza to the total population. We show the real distributions and the distributions from the apps GPS data. Two-sample KS test indicates that we cannot reject the hypothesis that the real and GPS sample come from the same distribution under the test, indicating lack of sampling bias under this test.

The two final tests of sampling bias are done on the age distribution and the gender distribution. We do not have direct access to the age and gender of the GPS users. However, an indirect test of sampling bias can be used using the patient data. We compare the distribution of age and gender in the full patient dataset with the distribution of age and gender of those patients that are localized in the GPS dataset. The hypothesis is that if the GPS dataset is biased by age or gender (for instance, if the app users over-represent younger people) then the distribution of localized patients in the GPS dataset should reflect this bias respect to the distribution of age and gender of the whole patient population. Figure B4 shows the respective PDF and CDF. We find that we cannot reject the hypothesis that the CDF of age from GPS data and the patient data come from the same distribution with p-value = 0.785 and KS distance = 0.039, indicating good coverage of age distribution (Figures B4a, b). The distributions of gender shown in Figures B5a, b indicate also the lack of bias in the gender distributions.

We conclude that the GPS sample does not have significant bias under

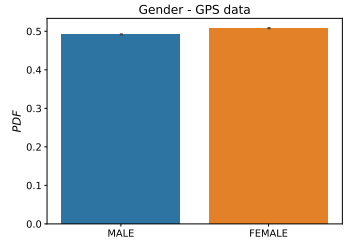


(a)

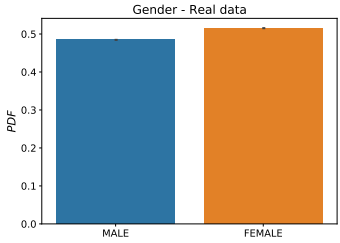


(b)

Figure B4: (a) PDF and **(b)** CDF of age distribution in the GPS geolocalized data compared with the real patient data. We cannot reject the hypothesis that both samples come from the same distribution under KS statistical testing.



(a)



(b)

Figure B5: (a) PDF and **(b)** CDF of gender distribution in the GPS geolocalized data compared with the real patient data suggesting lack of bias.

statistical testing in the considered demographic variables. The analyses above show that the distributions of location, economic status, age and gender are similar under two-sample KS test to the distributions of the real data. That is, we cannot reject the hypothesis that the real data and the GPS data comes from the same distribution. Thus, our results suggest that a collection of apps with GPS geolocalization provides a statistically significant sample to study the behaviour of the real population.

Regarding the use of global quantities, such as the betweeness centrality, it may not seen practical, in principle, due to the necessity to obtain the global network. However, according to the Covid Tracing Tracker from MIT Technology Review at `https://bit.ly/2Y1NMet`,

which tracks the contact tracing apps around the world, 35% of government-backed contact tracing apps are based on the same GPS technology used in our study and are able to provide the network of contacts needed for our algorithm to work. These GSP-based apps capture the necessary information on the global contact network needed to perform the present analysis, and our algorithm can be directly applied to them.

B.5 Robustness checks

Correlation between a unique mobile ID and a unique person. To test this correlation we investigated whether the geolocalized patients in the GPS datasets have (1) visited any hospital (near 200 m) or (2) have visited any pharmacy on the date of test results. We find that out of the 1,440 patients identified in the GPS dataset 1224 (85%) have been identified with the second test.

Persistence of centrality. It is of interest to discuss the persistence of centrality to see how stable is that set of central people. For instance, are the same people in the maximal k-core for all time or are we using a static network metric over the aggregated network? We have calculated the persistence of people in the different k-cores and find that they persist in the networks as a function of time. Results are shown in Figure B6.

Uncertainty in the first day of symptoms. The day of first symptoms is the date reported retrospectively by the patient of appearance of symptoms at the first consultation at a healthcare facility, thus, there is some uncertainty in its determination related to the patient's report of this date. This uncertainty could affect the results of the contact tracing. For proper functioning of the algorithms, data should be fed into the algorithm of contact tracing in real time, while for most of the cases the first date of symptoms is reported retroactively. In a mobile app implementation of contact tracing, the user should be given the capability to report the symptoms in real time as soon as they develop, via the app, thus diminishing uncertainties in the proper definition of the window of observation to detect contacts.

To verify the reliability of the protocol, we conduct a numerical study

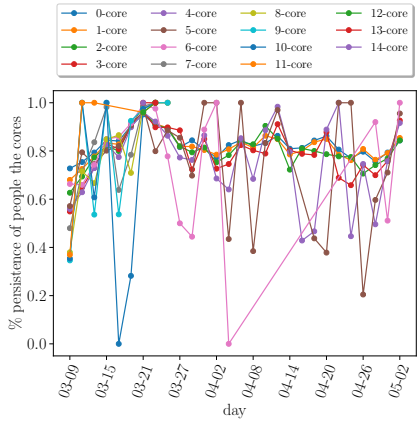


Figure B6: Persistence of people in the k-cores in the temporal networks. We plot the percentage of people in the cores from network to network. The persistance is calculated by the overlap of people in the k-shells from a time of observation to the next (three days later in this particular example).

with the random removal of infected nodes to mimic false positives on diagnosis. Figure B7 shows the relative error in the determination of the minimal number of people to quarantine as a function of false positives in the report of infected people. A false positive is an individual who reported to have symptoms but was not infected with COVID-19.

Temporal sampling The distribution of temporal sampling of GPS ping datapoints per user in our GPS dataset displays four peaks: around zero, at 5 minutes, 10 minutes and 20 minutes, see Figure B8. This is consistent with other apps using, e.g., the Google-Apple framework. Typically, this is a trade-off between accuracy and battery life. In our data, pinning distributions are uniform across day and night. We have investigated the robustness of our contact tracer to different ping intervals. The effect is particularly important since betweenness centrality and k-core are both macroscopic properties, meaning a small change in the network can create large changes everywhere in the network.

Figure 10b shows that the a minimal time interval that captures the correct behaviour in the probability to find a contact is around 30 min-

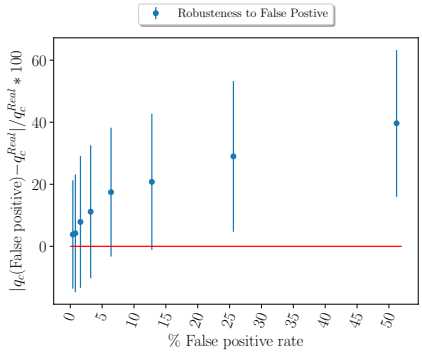


Figure B7: Normalized efficacy of BC centrality as a function of false positives in the report of infected people. A false positive is an individual who reported to have symptoms but was not infected with COVID-19. We plot the relative error in the determination of the minimal number of people to quarantine versus the false positive rate. The measure starts to deviate from linear behaviour beyond the error bars around 20% false positive rate.

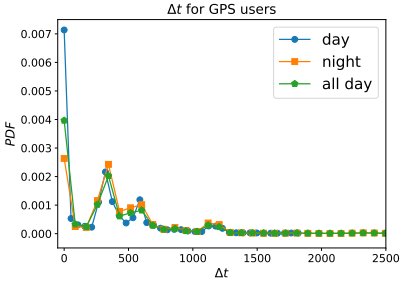


Figure B8: Distribution of the time interval between GPS pings during all day and separated by day and night.

utes. This is seen in Figure S10b where the correct decreasing behaviour of $< P_i[n] >_T$ with T appears after $T > 30$ minutes. Notice that this probability is supposed to decrease with increasing T. Considering T=30 minutes as a minimal time interval to find a contact, we would need at least two pings inside 30 minutes to properly define an interval of contacts. Therefore, we do not recommend to use longer ping intervals than

15 minutes, in order to have enough statistics to capture contact points.

Temporal features. Bias in the temporal features of the model. The periods of infectiousness and exposure are relaxed conservatively as explained in the text. We studied the robustness under the definition of these periods. We have relaxed the window forward in order to be symmetric with the relaxation of time before and after the symptoms by using 8 days after symptoms and find that the results do not change. This is expected since patients are expected to be either at home in quarantine or in hospital, in average, 5.5 days (95% CI 4.6 - 6.6 d) after first symptoms, according to [119]. Indeed, we find that the displacements of patients are highly reduced after the days of first symptoms. Thus, we expect that the majority of contacts are established before symptoms, highlighting the necessity of contact tracing back in time.

Unmatched cases. There are uncertainties in the number of cases matched between datasets due to the fact that only a fraction of patient cases can be matched to the GPS dataset. The unmatched cases cannot be matched at random to complete the data since these would ignore the correlations between the disease and behaviour. Thus, we do not consider the unmatched cases in the contact network. This unmatching is due to the incomplete coverage of the GPS dataset respect to the real population. However, we have checked in Section B.4 that the sampling coverage of the dataset is consistent with the real population which then minimizes the chances of small sampling bias. The spreading rate in our dataset is 0.112 as described in B.2. This value corresponds to $R_0^{\text{cal}} = 2.66$ consistent with the real R_0 value in the whole population.

Time-evolving weak ties. Weak ties in temporal networks have been investigated in [179]. In our case these weak ties are evolving with time as well. While we apply our definitions of centrality metrics to static networks defined over a week, we employ a moving window that calculates a new network every three days, thus extending the static definition of weak ties to temporal data.

Infected cases. We have based our modeling on the evolving number of new cases, which is not as robust estimate as the number of death cases. The main indicator in contact tracing is the date of first symp-

toms. In the absence of this datapoint, the day of hospitalization can be also used to estimate the date of first symptoms from the hospitalization data using the ensemble average of the time interval from first symptoms to hospitalization across the rest of the patients. While death is a more accurate metric than these two metrics, an estimate of the day of first symptoms is more difficult to obtained from the date of death, which anyways, occurs to only a fraction of the patients.

Contact tracing methods. In a contact tracing app based on GPS technology, a centralized server is needed where one entity has access to the GPS data. The algorithms discussed here are also feasible in contact tracing platforms using Bluetooth technology, as long as the detected neighbors are shared at a given point by a central server. This is because, the optimal tracing strategy requires the contacts of the contacts to build the network. On the other hand, the geolocalization of the red zone of contacts in a geographical map with precise locations requires the use of GPS data and cannot be performed with Bluetooth-based technology.

K-core infection. We mention an important point about the definition of the k-cores. Given a network, the same k-core can be composed of several disconnected components. This is what we see from Figures 17c and d, or for instance, the example 3-core in Figure 14c contains several disconnected components. Our analyses (displayed in Figures refsir-attackc and d) show that these disconnected components are connected by weak links, which if removed, may isolate the spreading of the virus inside one of the disconnected components of a k-core, containing the spreading of the pandemic to inside the disconnected component of the k-core. In other words, if a part of the k-core is infected, the disease will be controlled within a small group of that k-core and not extended to the rest of the network, if the affected component of the k-core is isolated from the rest by the removal of the weak link.

Model of infection. In our model, a transmission probability is calculated for a contact, which is then taken to be infectious if the probability is above a threshold. To avoid overestimating the importance of strong contacts we then define the recursive probability Eq. (3.5), $P_i[n]$, where one strong contact between two highly connected groups is not more im-

portant than many weak contacts. This regularization is highly efficient in converting the simple hit probability $p_i[n]$ which is observed to be not a good separator of a contact versus a non-contact since its distribution is not bimodal (orange curve) as observed in Figure 10a, into a bimodal distribution for $P_i[n]$ as observed in the figure (blue curve). Thus, $P_i[n]$, which takes into account the importance of many small contacts separates well the contacts, and does not need of a precise threshold to be implemented. In fact, any threshold in the range $0.1 < p_c < 0.9$ gives similar results in the contact model since most of the cases are concentrated in the extreme cases $P_i[n] \approx 1$ and $P_i[n] \ll 1$ or zero. Thus, while the thresholding of $p_i[n]$ has a danger of throwing out many weak links, the use of $P_i[n]$ regularize this function in the proper way such that the effect of choosing a threshold is minimized. That is, since the distribution of $P(P_i[n])$ (Figure 10a) is highly bimodal concentrating near zero and one, then we were able to separate well a contact $P_i[n] \approx 1$ from a non-contact $P_i[n] = 0$, and the need for a threshold to distinguish between these two extremes disappears, since the probability is nearly zero between these two extreme values as shown in the figure.

Dynamics of weak links. In principle, the networks are calculated continuously in time, and therefore the weak links are identified as they form and break k-cores. If by removing an individual, somebody else replaces that role, then a new network should identify this new individual added as a new weak link. However, this process may become somehow impractical due to the continuous removal of those weak links. This problem can be treated by determining the roles or occupations of the weak links and then search for a way to remove this risk by targeting those roles or occupations. This can be achieved by the analysis shown in, for instance, Figures 12c and d, by geolocalizing the weak links and k-core in the map to discover the occupations/roles or places visited by the weak links and develop a targeted approach accordingly. For instance, Figure 12d shows where these occupations and roles contributing to the transmission chain are occurring. Targeting these occupations and places to remove the risk is an efficient way to break the chain of transmission.

Uncertainties in infected population. We match the infected indi-

vidual by a rule that uses the geolocation of the individual at the address of the patient. This method may cause some uncertainty in the results as some fraction of the infected individuals could be in principle replaced by non-infected individuals. To study this problem, we test that the identified mobile ID has also visited the hospital and or a pharmacy on the date of test results. Using this second test, we corroborated the matching between patients and mobile users.

Global measures and scalability. A drawback for the use of a global measure like betweeness centrality is the poor scalability of the measure for large system sizes. However, there are linear approximations of these algorithms that can be used to approximate the metrics for large systems [123]. For larger datasets approximate fast algorithms can be used to calculated BC [123]. Furthermore, once the network is obtained, the chain of transmission can be destroyed by other measures other than BC, which scales linearly with system size and are quite fast to calculate like the degree or CI, although not as optimal as through the weak links as shown by our results.

Asymptomatic cases. Detecting asymptomatic cases is one of the biggest challenge of the COVID-19 pandemic. We have included the existence of contacts with asymptomatic in our model. Our method allows, in principle, the determination of possible contacts with asymptomatic infectious people. As explained in Figure 9a and further explained in Figure B1, we extend the exposure period to -14 days from the day of first symptoms. At the same time, the infectious period starts -2 days from symptoms according to [117]. Therefore the contacts identified between -2 days and -14 days (labeled as E′ in Figure 9a and also as the inbound contacts in Figure B1) correspond to inbound exposures from asymptomatic carriers. Thus, our model treats asymptomatic cases by considering this exposure period and accounts for possible two-chains of infection as shown in Figure B1. Since the exposure period with asymptomatic (-14 to -2 days) is longer than the period of infectiousness (-2 to +5 days), we obtain a larger number of asymptomatic exposures than infectious transmission contacts per patient.

Detecting asymptomatic is the key to stop this pandemic. Asymp-

tomatic detected at exposure E′ should be immediately tested, even though they have not presented symptoms. Beyond our modeling, we are not aware of other contact methods that have attempted to detect asymptomatic contacts (beyond large-scale widespread testing). Our contact tracing algorithms might be able to detect those asymptomatic transmissions which are critical to stop the pandemic.

Quantification of uncertainty in data. Quantification of uncertainty in data plotted in the figures is done via the calculation of the standard error (SD). For the SIR models, the errors are computed as the SE for a given value of the k-core. We notice that some plots are single instances, like for instance the giant connected component, and do not have SE.

B.6 Privacy considerations

Ethics and privacy consideration of data sharing and using large datasets in biomedical research.

The patient datasets were collected by the Epidemiological Surveillance Department, Fortaleza Health Secretariat (CEVEPI) at the Prefeitura de Fortaleza, Ceará, Brazil by the team of Dr. Antonio S. Lima Neto. The present project follows the recommendations of the Wellcome Trust 2016 *"Statement on data sharing in public health emergencies"*, that states: *"In the context of a public health emergency of international concern, there is an imperative on all parties to make any information available that might have value in combating the crisis"* and the statement of the WHO: *"data are the basis for all sound public health actions"*. The Epidemiological Surveillance Department have provided the anonymized data on COVID-19 patients in the City of Fortaleza including the SARS-COV-2 test detection date and first day of symptoms of COVID-19 geocoded cases. Prior to be analyzed, the dataset was completely de-identified at CEVEPI. The data did not include any element that allow us to make a full profile of the patient, such as, the name of the patient, neither the residential address. Additionally, there were no codes associated with the variables that may allow individuals to be identified.

The policy statement for the use of the data states that the information

was collected and anonymized by CEVEPI and analyzed by the team at CCNY and UFC, and then deleted completely from the servers of the team at both CCNY and UFC. At the time of submission, the provided datasets have already been deleted from the servers at CCNY and UFC. The data were made available to the team over a period from March 1, 2020 to June 7, 2020 for the purpose of helping the government contain the pandemic. The protocol of contact tracing developed in this study was done within the framework of the public functions of the government aimed at protecting and guaranteeing the public health of the citizens.

The team has signed confidentially agreements stating that the member of the team are responsible for the custody of the data received and that we guarantee the privacy, the confidentiality, anonymization and used of the obtained information of the patients and any third party. The confidentially agreement also states that the team will receive the data only for the purpose of the current investigation and that all data will be erased from the servers at CCNY and UFC at the end of the project in July 2020. All data has been already deleted from the servers at CCNY and UFC and remains under the control of Department of Epidemiological Surveillance, Fortaleza Health Secretariat.

Given the extraordinary nature of this health emergency, it is critical for governments and agencies to share these data with scientists to execute critical studies under the understanding that by using the contact tracing algorithms developed in this study, we could help avoid more casualties during the pandemic. The patient dataset from the Health Department authorities in Fortaleza has the required ethics approval from the IRB at CCNY and UFC. Consent to use these data was given by the Mayor of Fortaleza and the Prefeitura of the City of Fortaleza. The original dataset is kept under the care of the Health Secretariat of Fortaleza.

Appendix C

Daily statistics, hashtags and rescaling procedure

Here we investigate the daily statistics for the data collection, as well as the individuation of bots in the dataset. We also give more details about the rescaling procedure, introduced in Chapter 4, and the effect of T_0 on the final predictions.

C.1 Tweets and Users classification

Tables C1 and C2 show the statistics of the collection, in terms of tweets and users, respectively. We can distinguish two columns: **Bots** and **Users**. In the case of tweets, bots refers to the number of tweets posted by a bot, while users refer to the tweets posted by a genuine account. Table C3 reports the top 20 hashtags between March and July, 2019. FF indicates the hashtags is in favor of the Fernández-Fernández formula while MP refers to the supporters of the Macri-Pichetto formula. Figure C1 shows the average predictions of the top 5 trusted polls, i.e., Real Time Data, Management & Fit, Opinaia, Giacobbe and Elypsis (thin lines) and the the AI-model window average (tick lines).

Tweets	Bots	Users
Total	538359	67336507
Daily	2243	280568
Daily classified	1617	211229
Daily classified MP	619	114653
Daily classified FF	998	96576

Table C1: Tweets statistics. We report the Total number of tweets collected. The average daily number of tweets classified (Daily classified) and the average daily classified tweets for each candidate.

Tweets	Bots	Users
Total	17953	2252551
Daily	732	83330
Daily classified	560	63808
Daily classified MP	198	31497
Daily classified FF	362	32310

Table C2: Users statistics. We report the Total number of users collected. The average daily number of users classified (Daily classified) and the average daily classified users for each candidate.

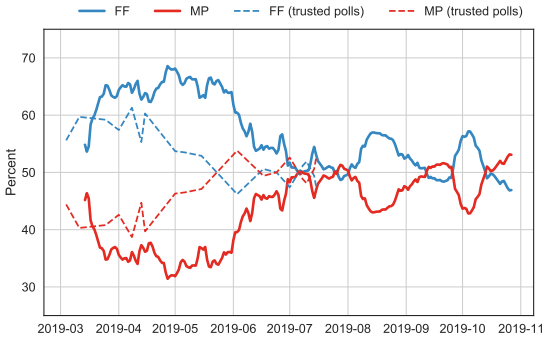


Figure C1: Instantaneous prediction compared with trusted polls. Thick lines represent AI predictions and dashed lines represent the average trusted polls.

#	Hashtag	Camp	Count
1	sisepuede	MP	216757
2	macri	MP	102172
3	albertopresidente	FF	65644
4	axelgobernador	FF	60920
5	juntosporelcambio	MP	52617
6	yovotomm	MP	49557
7	ladamosvuelta	MP	47546
8	cfk	FF	38718
9	cambiemos	MP	38538
10	sevan	FF	32110
11	habraconsecuencias	FF	28406
12	24a	MP	23819
13	macrihacetecrgo	FF	21364
14	novuelvenmas	MP	21054
15	cronicaanunciada	FF	20293
16	frentedetodos	FF	18115
17	albertoycristina	FF	17482
18	sinceramente	FF	15403
19	juntossomosimparables	MP	14405
20	sevanenprimeravuelta	FF	14091

Table C3: The top 20 hashtags from March and July in 2019. The camp field represents the classification: MP stays for Macri and FF for Fernández (from the name of the running mate Cristina Fernández de Kirchner. Count indicates the number of time a given hashtag appears in the dataset.

C.2 Rescaling Method

Of common practice in the traditional pollster methods, this method allows to rescale the Twitter population, Figure reffig:twitter-population to the Census data in Figure 20c. Let $N(c, g, a)$, with $c \in \{M, F, T\}$ be the number of voters supporting one of the candidate, with age a (a = [16, 30], [31, 50], [51, 65] or [66, ∞]). and gender g (g = female or male). The percentage of voters p supporting a given candidate c is

$$p_c = \sum \frac{N(c, g, a)}{N(g, a)} * \frac{N_{Census}(g, a)}{N_{Census}} \tag{C.1}$$

with $N(g, a)$ total number of user if the dataset with a gender g and age a, $N_{Census}(g, a)$ total number of citizens with a gender g and age a and N_{Census} total number of citizens in the Census data. p_c is always between 1 and 0 and can be directly compared with the pollster prediction.

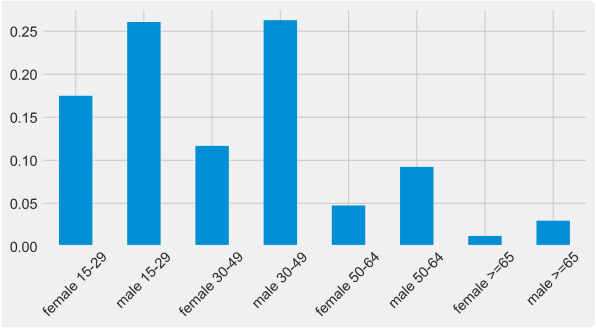


Figure C2: Distribution of Twitter populations by age and gender during 2019 Argentina election.

C.3 Effect of the initial time on the final result

As discussed in the main text, different values for T_0 can lead to different results. If so, the problem of finding a good substitute to traditional polls would not be solved, but perhaps moved to another problem such as find the best value for T_0, defined as the value that approximate the official

results of the elections. To investigate this effect, we have recalculated the cumulative average by changing the origin of measurement T_0 in Figure C3. Predictions for the general elections cluster around the same value. We use the fluctuations at a given time t to compute the error associate to the prediction at the same time. We define the error as the standard deviation over the results of different realizations with $t < T_0$. Regarding MODEL 3, the estimated average error is 0.53%.

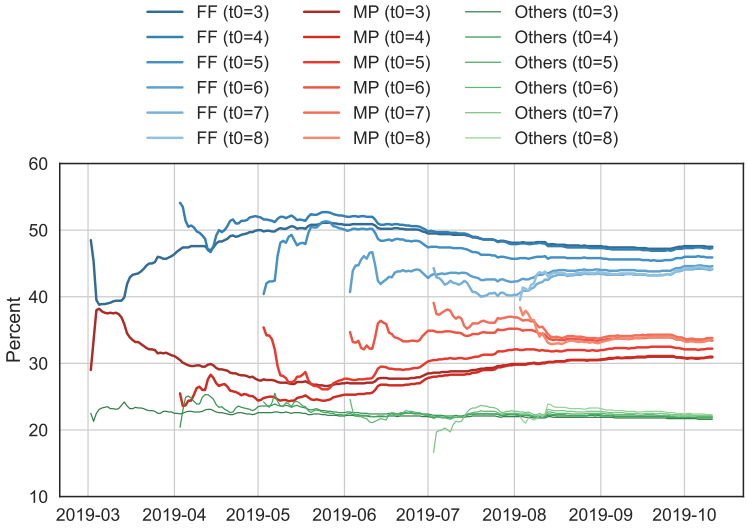


Figure C3: Cumulative prediction for different initial times t_0 (March 3, April 4, etc.). While we consider different t_0, the predictions at the day of the election clusters well around the results of the PASO.

Bibliography

[1] US inflation calculator. *US inflation calculator*.

[2] The Guardian. *In just 25 years, the mobile phone has transformed the way we communicate*. 2010. URL: https://www.theguardian.com/business/2010/jan/01/25-years-phones-transform-communication.

[3] Guido Caldarelli. "A perspective on complexity and networks science". In: *Journal of Physics: Complexity* 1.2 (2020), p. 021001.

[4] Peter T Saunders and Mae-Wan Ho. "On the increase in complexity in evolution". In: *Journal of Theoretical Biology* 63.2 (1976), pp. 375–384.

[5] Philip W Anderson. "More is different". In: *Science* 177.4047 (1972), pp. 393–396.

[6] The Guardian. *Study: less than 1% of the world's data is analysed, over 80% is unprotected*. 2012. URL: https://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume.

[7] IBM. *IBM Annual Report*. 2017. URL: https://www.ibm.com/annualreport/assets/past-reports/2017-ibm-annual-report.pdf.

[8] Bernard Marr. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. 2019. URL: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/.

[9] Béla Bollobás. *Random Graphs*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001. DOI: 10.1017/CBO9780511814068.

[10] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509.

[11] Per Bak, Chao Tang, and Kurt Wiesenfeld. "Self-organized criticality". In: *Physical review A* 38.1 (1988), p. 364.

[12] Nigel Goldenfeld and Leo P Kadanoff. "Simple lessons from complexity". In: *science* 284.5411 (1999), pp. 87–89.

[13] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *nature* 393.6684 (1998), p. 440. DOI: 10.1038/30918.

[14] Petter Holme. "Rare and everywhere: Perspectives on scale-free networks". In: *Nature Communications* 10 (1 2007), p. 1016. DOI: 10.1038/s41467-019-09038-8.

[15] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[16] Guido Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.

[17] G. Cimini et al. "The Statistical Physics of Real-World Networks". In: *Nature Reviews Physics* 1 (1 2019), pp. 58–71. DOI: 10.1038/s42254-018-0002-6.

[18] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: *Reviews of modern physics* 87.3 (2015), p. 925.

[19] Erica Klarreich. "Scant evidence of power laws found in real-world networks". In: *Quanta Magazine. Feb* 15 (2018).

[20] Matteo Serafino et al. "True scale-free networks hidden by finite size effects". In: *Proceedings of the National Academy of Sciences* 118.2 (2021).

[21] A. Barrat et al. "The architecture of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 101.11 (2004), pp. 3747–3752. DOI: 10.1073/pnas.0400087101.

[22] Bin Zhou, Xiangyi Meng, and H. Eugene Stanley. "Power-law distribution of degree-degree distance: A better representation of the scale-free property of complex networks". In: *Proceedings of the National Academy of Sciences* (2020). DOI: 10.1073/pnas.1918901117. eprint: https://www.pnas.org/content/early/2020/06/12/1918901117.full.pdf.

[23] Alessandro Vespignani. "Predicting the behavior of techno-social systems". In: *Science* 325.5939 (2009), pp. 425–428.

[24] Alessandro Vespignani. "Modelling dynamical processes in complex socio-technical systems". In: *Nature physics* 8.1 (2012), pp. 32–39.

[25] Herbert W Hethcote and James A Yorke. *Gonorrhea transmission dynamics and control*. Vol. 56. Springer, 2014.

[26] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[27] David MJ Lazer et al. "The science of fake news". In: *Science* 359.6380 (2018), pp. 1094–1096.

[28] Alexandre Bovet and Hernán A Makse. "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature communications* 10.1 (2019), pp. 1–14.

[29] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[30] Guido Caldarelli et al. "The role of bot squads in the political propaganda on Twitter". In: *Communications Physics* 3.1 (2020), pp. 1–15.

[31] Elaine Reddy. *How the USGS uses Twitter data to track earthquakes*. 2015. URL: https://blog.twitter.com/en_us/a/2015/usgs-twitter-data-earthquake-detection.html.

[32] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[33] Hamed Jelodar et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11 (2019), pp. 15169–15211.

[34] Bing Liu et al. "Sentiment analysis and subjectivity." In: *Handbook of natural language processing* 2.2010 (2010), pp. 627–666.

[35] Anna D. Broido and Aaron Clauset. "Scale-free networks are rare". In: *Nature Communications* 10 (1 2019), p. 1017. DOI: 10.1038/s41467-019-08746-5.

[36] M E Fisher. "The theory of equilibrium critical phenomena". In: *Reports on Progress in Physics* 30.2 (1967), p. 615. DOI: 10.1088/0034-4885/30/2/306.

[37] H. Eugene Stanley. *Introduction to phase transitions and critical phenomena*. Oxford University Press, 1971.

[38] K Binder and D Heermann. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Springer-Verlag, 1992.

[39] Jae-Kwon Kim, Adauto J.F. De Souza, and D. P. Landau. "Numerical computation of finite size scaling functions: An alternative approach to finite size scaling". In: *Physical Review E* 54.3 (1996), pp. 2291–2297. DOI: 10.1103/PhysRevE.54.2291.

[40] H. Eugene Stanley. "Scaling, universality, and renormalization: Three pillars of modern critical phenomena". In: *Reviews of Modern Physics* 71 (2 1999), S358–S366. DOI: 10.1103/RevModPhys.71.S358.

[41] Sidney Redner. *A guide to first-passage processes*. Cambridge University Press, 2001.

[42] Álvaro Corral, Rosalba Garcia-Millan, and Francesc Font-Clos. "Exact derivation of a finite-size scaling law and corrections to scaling in the geometric Galton-Watson process". In: *PLoS ONE* 11.9 (2016), e0161586. DOI: 10.1371/journal.pone.0161586.

[43] Romualdo Pastor-Satorras and Alessandro Vespignani. "Immunization of complex networks". In: *Physical review E* 65.3 (2002), p. 036104.

[44] Mark EJ Newman. "Spread of epidemic disease on networks". In: *Physical review E* 66.1 (2002), p. 016128.

[45] Matteo Serafino et al. "Superspreading k-cores at the center of COVID-19 pandemic persistence". In: *arXiv preprint arXiv:2103.08685* (2021).

[46] *Grandata-United Nations Development Programme partnership to combat COVID-19 with data*. 2020. URL: https://covid.grandata.com (visited on 03/05/2021).

[47] Sergey N Dorogovtsev, Alexander V Goltsev, and Jose Ferreira F Mendes. “K-core organization of complex networks”. In: *Physical review letters* 96.4 (2006), p. 040601.

[48] Shai Carmi et al. “A model of Internet topology using k-shell decomposition”. In: *Proceedings of the National Academy of Sciences* 104.27 (2007), pp. 11150–11154.

[49] J Ignacio Alvarez-Hamelin et al. “Large scale networks fingerprinting and visualization using the k-core decomposition”. In: *Advances in neural information processing systems*. 2006, pp. 41–50.

[50] Maksim Kitsak et al. “Identification of influential spreaders in complex networks”. In: *Nature physics* 6.11 (2010), pp. 888–893.

[51] Mark S Granovetter. “The strength of weak ties”. In: *American journal of sociology* 78.6 (1973), pp. 1360–1380.

[52] Linton C Freeman. “Centrality in social networks conceptual clarification”. In: *Social networks* 1.3 (1978), pp. 215–239.

[53] Noah E Friedkin. “Theoretical foundations for centrality measures”. In: *American journal of Sociology* 96.6 (1991), pp. 1478–1504.

[54] Zhenkun Zhou et al. “Why polls fail to predict elections”. In: *arXiv preprint arXiv:2101.11389* (2021).

[55] Michael P. H. Stumpf and Mason A. Porter. “Critical truths about power laws”. In: *Science* 335.6069 (2012), pp. 665–666. DOI: 10.1126/science.1216142.

[56] Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. “Large-scale analysis of Zipf’s law in English texts”. In: *PloS one* 11.1 (2016).

[57] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111.

[58] Ivan Voitalov et al. “Scale-free networks well done”. In: *Physical Review Research* 1 (3 2019), p. 033034. DOI: 10.1103/PhysRevResearch.1.033034.

[59] Martin Gerlach and Eduardo G. Altmann. “Testing Statistical Laws in Complex Systems”. In: *Physical Review Letters* 122 (16 2019), p. 168301. DOI: 10.1103/PhysRevLett.122.168301.

[60] Mark Newman. *Autonomous systems (DIMACS10)*. http://konect.cc/networks/dimacs10-as-22july06/. 2006.

[61] Pierre-Gilles De Gennes and Pierre-Gilles Gennes. *Scaling concepts in polymer physics*. Cornell university press, 1979.

[62] Samuel Safran. *Statistical thermodynamics of surfaces, interfaces, and membranes*. CRC Press, 2018.

[63] David R Nelson, Tsvi Piran, and Steven Weinberg. *Statistical mechanics of membranes and surfaces*. World Scientific, 2004.

[64] Per Bak, Chao Tang, and Kurt Wiesenfeld. "Self-organized criticality: An explanation of the 1/f noise". In: *Physical Review Letters* 59.4 (1987), pp. 381–384. DOI: 10.1103/PhysRevLett.59.381.

[65] P L Krapivsky and S Redner. "Finiteness and fluctuations in growing networks". In: *Journal of Physics A: Mathematical and General* 35.45 (2002), pp. 9517–9534. DOI: 10.1088/0305-4470/35/45/302.

[66] Somendra M Bhattacharjee and Flavio Seno. "A measure of data collapse for scaling". In: *Journal of Physics A: Mathematical and General* 34.33 (2001), p. 6375. DOI: 10.1088/0305-4470/34/33/302.

[67] Jérôme Houdayer and Alexander K Hartmann. "Low-temperature behavior of two-dimensional Gaussian Ising spin glasses". In: *Physical Review B* 70.1 (2004), p. 014418. DOI: 10.1103/PhysRevB.70.014418.

[68] Oliver Melchert. *autoScale.py–A program for automatic finite-size scaling analyses: A user's guide*. https://arxiv.org/abs/0910.5403. 2009.

[69] Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. "Subnets of scale-free networks are not scale-free: Sampling properties of networks". In: *Proceedings of the National Academy of Sciences* 102.12 (2005), pp. 4221–4224. DOI: 10.1073/pnas.0501179102.

[70] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. "Statistical properties of sampled networks". In: *Physical Review E* 73 (1 2006), p. 016102. DOI: 10.1103/PhysRevE.73.016102.

[71] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. "powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions". In: *PLoS ONE* 9.1 (2014), e85777. DOI: 10.1371/journal.pone.0085777.

[72] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. "Structure of growing networks with preferential linking". In: *Physical Review Letters* 85.21 (2000), pp. 4633–4636. DOI: 10.1103/PhysRevLett.85.4633.

[73] Paul Erdös and Alfred Rényi. "On random graphs". In: *Publicationes Mathematicae Debrecen* 6 (1959), pp. 290–297.

[74] Fabien Viger and Matthieu Latapy. "Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence". In: *Computing and Combinatorics*. Ed. by Lusheng Wang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 440–449. ISBN: 978-3-540-31806-4.

[75] Ulrich Stelzl et al. "A human protein-protein interaction network: a resource for annotating the proteome". In: *Cell* 122.6 (2005), pp. 957–968.

[76] Ron Milo et al. "Superfamilies of evolved and designed networks". In: *Science* 303.5663 (2004), pp. 1538–1542. DOI: 10.1126/science.1089167.

[77] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graph Evolution: Densification and Shrinking Diameters". In: *ACM Trans. Knowl. Discov. Data* 1.1 (2007). ISSN: 1556-4681. DOI: 10.1145/1217299.1217301.

[78] Michael Ley. "The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives". In: *String Processing and Information Retrieval*. Ed. by Alberto H. F. Laender and Arlindo L. Oliveira. Springer Berlin Heidelberg, 2002, pp. 1–10. ISBN: 978-3-540-45735-0.

[79] S. N. Dorogovtsev and J. F. F. Mendes. "Evolution of networks". In: *Advances in Physics* 51.4 (2002), pp. 1079–1187. DOI: 10.1080/00018730110112519.

[80] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. "Cut-offs and finite size effects in scale-free networks". In: *The European Physical Journal B* 38.2 (2004), pp. 205–209. DOI: 10.1140/epjb/e2004-00038-8.

[81] William Aiello, Fan Chung, and Linyuan Lu. "A Random Graph Model for Power Law Graphs". In: *Experimental Mathematics* 10.1 (2001), pp. 53–66. DOI: 10.1080/10586458.2001.10504428.

[82] Yong Wang, Xiaochun Yun, and Yifei Li. "Analyzing the characteristics of Gnutella overlays". In: *Fourth International Conference on Information Technology (ITNG'07)*. IEEE. 2007, pp. 1095–1100. DOI: 10.1109/ITNG.2007.39.

[83] M. Gjoka et al. "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs". In: *2010 Proceedings IEEE INFOCOM*. 2010, pp. 1–9. DOI: 10.1109/INFCOM.2010.5462078.

[84] Johan Ugander et al. *The anatomy of the Facebook social graph*. http://arxiv.org/abs/1111.4503. 2011.

[85] Tao Zhou et al. "Emergence of scale-free leadership structure in social recommender systems". In: *PLoS ONE* 6.7 (2011), e20648. DOI: 10.1371/journal.pone.0020648.

[86] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the Internet topology". In: *ACM SIGCOMM Computer Communication Review*. Vol. 29. ACM Press, 1999, pp. 251–262. DOI: 10.1145/316188.316229.

[87] Lada A. Adamic and Bernardo A. Huberman. "Power-law distribution of the World Wide Web". In: *Science* 287.5461 (2000), p. 2115. DOI: 10.1126/science.287.5461.2115a.

[88] Guido Caldarelli, Riccardo Marchetti, and Luciano Pietronero. "The fractal properties of Internet". In: *Europhysics Letters* 52.4 (2000), pp. 386–391. DOI: doi:10.1209/epl/i2000-00450-8.

[89] Kwang-Il Goh et al. "Classification of scale-free networks". In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12583–12588. DOI: 10.1073/pnas.202301299.

[90] Alan Mislove et al. "Measurement and Analysis of Online Social Networks". In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. IMC '07. New York, NY, USA: ACM, 2007, pp. 29–42. DOI: 10.1145/1298306.1298311.

[91] G. Bianconi and A.L. Barabási. "Bose-Einstein condensation in complex network". In: *Physical Review Letters* 86 (2001), pp. 5632–5635. DOI: 10.1103/PhysRevLett.86.5632.

[92] G. Caldarelli et al. "Scale-free networks from varying vertex intrinsic fitness". In: *Physical Review Letters* 89.25 (2002), p. 258702. DOI: 10.1103/PhysRevLett.89.258702.

[93] M. Medo, G. Cimini, and S. Gualdi. "Temporal effects in the growth of networks". In: *Physical Review Letters* 107.23 (2011), p. 238701. DOI: 10.1103/PhysRevLett.107.238701.

[94] Michael Mitzenmacher. "A brief history of generative models for power-law and log-normal". In: *Internet Mathematics* 1 (2004), pp. 226–251. DOI: 10.1080/15427951.2004.10129088.

[95] Reiko Tanaka. "Scale-rich metabolic networks". In: *Physical Review Letters* 94 (16 2005), p. 168101. DOI: 10.1103/PhysRevLett.94.168101.

[96] Chaoming Song, Shlomo Havlin, and Hernán A Makse. "Origins of fractality in the growth of complex networks". In: *Nature physics* 2.4 (2006), pp. 275–281.

[97] Fragkiskos Papadopoulos et al. "Popularity versus similarity in growing networks". In: *Nature* 489.7417 (2012), pp. 537–540.

[98] Guillermo García-Pérez, Marián Boguñá, and M. Ángeles Serrano. "Multiscale unfolding of real networks by geometric renormalization". In: *Nature Physics* 14 (6 2018), pp. 583–589. DOI: 10.1038/s41567-018-0072-5.

[99] Chaoming Song, Shlomo Havlin, and Hernan A Makse. "Self-similarity of complex networks". In: *Nature* 433.7024 (2005), p. 392. DOI: 10.1038/nature03248.

[100] J S Kim et al. "Fractality and self-similarity in scale-free networks". In: *New Journal of Physics* 9.6 (2007), p. 177. DOI: 10.1088/1367-2630/9/6/177.

[101] World Health Organization. "Coronavirus disease (COVID-19) Situation Report - 124". In: (2020).

[102] Ruiyun Li et al. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)". In: *Science* 368.6490 (2020), pp. 489–493.

[103] Matteo Chinazzi et al. "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak". In: *Science* 368.6489 (2020), pp. 395–400.

[104] P Howell O'Neill, Tate Ryan-Mosley, and Bobbie Johnson. "A flood of coronavirus apps are tracking us. Now it's time to keep track of them". In: *MIT Technology Review* (2020).

[105] Jennifer Valentino-DeVries, Natasha Singer, and Aaron Krolik. "A scramble for virus apps that do no harm". In: *The New York Times* (2020).

[106] Singapore Government. *TraceTogether, safer together*. 2020. URL: https://www.tracetogether.gov.sg/.

[107] Israel Ministry of Health. *The Ministry of Health App for Fighting the Spread of Coronavirus*. 2020. URL: https://bit.ly/3gUMxGC (visited on 03/05/2021).

[108] Helsenorge. *Together we can fight coronavirus*. 2020. URL: https://helsenorge.no/coronavirus/smittestopp (visited on 03/05/2021).

[109] K. (South Dakota Department of Health) Governor Noem. *Care 19 App.* 2020. URL: https://covid.sd.gov/care19app.aspx (visited on 03/05/2021).

[110] PH O'Neill. "Apple and Google are building coronavirus tracking into iOS and Android". In: *Acesso em* 30 (2020).

[111] Iceland COVID-19. *Contagion tracing is a community affair*. 2020. URL: https://bit.ly/2Y0gtZ0 (visited on 03/05/2021).

[112] Luca Ferretti et al. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing". In: *Science* 368.6491 (2020).

[113] Nuria Oliver et al. *Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle*. 2020.

[114] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". In: *nature* 406.6794 (2000), pp. 378–382.

[115] Reuven Cohen et al. "Breakdown of the internet under intentional attack". In: *Physical review letters* 86.16 (2001), p. 3682.

[116] Flaviano Morone and Hernán A Makse. "Influence maximization in complex networks through optimal percolation". In: *Nature* 524.7563 (2015), pp. 65–68.

[117] Xi He et al. "Temporal dynamics in viral shedding and transmissibility of COVID-19". In: *Nature medicine* 26.5 (2020), pp. 672–675.

[118] Qun Li et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia". In: *New England journal of medicine* (2020).

[119] Steven Sanche et al. “High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2”. In: *Emerging infectious diseases* 26.7 (2020), pp. 1470–1477.

[120] Alessandro Vespignani and Guido Caldarelli. *Large Scale Structure and Dynamics of Complex Networks: From information technology to finance and natural science*. Vol. 2. World scientific, 2007.

[121] Marc Barthelemy. “Betweenness centrality in large complex networks”. In: *The European physical journal B* 38.2 (2004), pp. 163–168.

[122] Laurent Hébert-Dufresne et al. “Global efficiency of local immunization on complex networks”. In: *Scientific reports* 3.1 (2013), pp. 1–8.

[123] Kamesh Madduri et al. “A faster parallel algorithm and efficient multithreaded implementations for evaluating betweenness centrality on massive datasets”. In: *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE. 2009, pp. 1–8.

[124] Sen Pei and Hernán A Makse. “Spreading dynamics in complex networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.12 (2013), P12002.

[125] Flaviano Morone, Gino Del Ferraro, and Hernán A Makse. “The k-core as a predictor of structural collapse in mutualistic ecosystems”. In: *Nature physics* 15.1 (2019), pp. 95–102.

[126] Alison P Galvani and Robert M May. “Dimensions of superspreading”. In: *Nature* 438.7066 (2005), pp. 293–295.

[127] James O Lloyd-Smith et al. “Superspreading and the effect of individual variation on disease emergence”. In: *Nature* 438.7066 (2005), pp. 355–359.

[128] Thomas R Frieden and Christopher T Lee. “Identifying and interrupting superspreading events—implications for control of severe acute respiratory syndrome coronavirus 2”. In: *Emerging infectious diseases* 26.6 (2020), p. 1059.

[129] Marcel Salathé and James H Jones. “Dynamics and control of diseases in networks with community structure”. In: *PLoS Comput Biol* 6.4 (2010), e1000736.

[130] World Health Organization. “‘Managing and monitoring contacts daily”. In: *Contact tracing in the context of COVID-19* (2020).

[131] Roger Tourangeau, Frederick G Conrad, and Mick P Couper. *The science of web surveys*. Oxford University Press, 2013.

[132] Claire Durand and André Blais. "Quebec 2018: A failure of the polls?" In: *Canadian Journal of Political Science/Revue canadienne de science politique* 53.1 (2020), pp. 133–150.

[133] Abebaye Bekele. "Are faulty opinion polls to blame for Brexit?" In: (2018).

[134] The Economist. "A new iPhone feature poses a threat to opinion pollsters". In: (2019). URL: https://www.economist.com/united-states/2019/09/26/a-new-iphone-feature-poses-a-threat-to-opinion-pollsters.

[135] J. Jacobs and B. House. "Trump Says He Expected to Lose Election Because of Poll Results". In: (2019). URL: https://www.bloomberg.com/politics/articles/2016-12-14/trump-says-he-expected-to-lose-election-because-of-poll-results.

[136] Courtney Kennedy et al. *"An Evaluation of 2016 Election Polls in the US" American Association for Public Opinion Research*. 2017.

[137] C. Kennedy and H. Hartig. "Response rates in telephone surveys have resumed their decline". In: (2019). URL: https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline.

[138] J Gregory Payne. "The Bradley effect: Mediated reality of race and politics in the 2008 US presidential election". In: *American Behavioral Scientist* 54.4 (2010), pp. 417–435.

[139] Kokil Jaidka et al. "Predicting elections from social media: a three-country, three-method comparative study". In: *Asian Journal of Communication* 29.3 (2019), pp. 252–273.

[140] Andreas Jungherr. "Twitter use in election campaigns: A systematic literature review". In: *Journal of information technology & politics* 13.1 (2016), pp. 72–91.

[141] Alexandre Bovet, Flaviano Morone, and Hernán A Makse. "Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump". In: *Scientific reports* 8.1 (2018), pp. 1–16.

[142] Andranik Tumasjan et al. “Predicting elections with twitter: What 140 characters reveal about political sentiment”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 4. 1. 2010.

[143] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. “Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im “predicting elections with twitter: What 140 characters reveal about political sentiment””. In: *Social science computer review* 30.2 (2012), pp. 229–234.

[144] Manish Gaurav et al. “Leveraging candidate popularity on Twitter to predict election outcome”. In: *Proceedings of the 7th workshop on social network mining and analysis*. 2013, pp. 1–8.

[145] Catherine Lui, P Takis Metaxas, and Eni Mustafaraj. “On the predictability of the US elections through search volume activity”. In: (2011).

[146] Adam Bermingham and Alan Smeaton. “On using Twitter to monitor political sentiment and predict election results”. In: *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*. 2011, pp. 2–10.

[147] Andrea Ceron et al. “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France”. In: *New media & society* 16.2 (2014), pp. 340–358.

[148] Guido Caldarelli et al. “A multi-level geographical study of Italian political elections from Twitter data”. In: *PloS one* 9.5 (2014), e95809.

[149] Prabhsimran Singh, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. “Forecasting the 2016 US presidential elections using sentiment analysis”. In: *Conference on e-Business, e-Services and e-Society*. Springer. 2017, pp. 412–423.

[150] Prabhsimran Singh et al. “Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections”. In: *Government Information Quarterly* 37.2 (2020), p. 101444.

[151] Mark Newman. *Networks*. Oxford University Press, 2018.

[152] Alfredo Cuzzocrea et al. “Edge betweenness centrality: A novel algorithm for QoS-based topology control over wireless sensor networks”. In: *Journal of Network and Computer Applications* 35.4 (2012), pp. 1210–1217.

[153] Leticia Bode and Kajsa E Dalrymple. “Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter”. In: *Journal of Political Marketing* 15.4 (2016), pp. 311–332.

[154] F. Morone A. Bovet S. Pei and H. A. Makse. *The Science of Influencers Using Mathematically Rigorous Theories - Understanding the Future of Society, Biology, Markets and Ecosystems*. Springer Nature, Switzerland, 2019.

[155] Barbara R Jasny and Richard Stone. *Prediction and its limits*. 2017.

[156] Wikipedia. *2019 Argentine general election*. URL: https://en.wikipedia.org/wiki/2019_Argentine_general_election#Opinion_polls.

[157] Wikipedia. *Compilation of Argentina election polls for PASO and general elections*. URL: https://es.wikipedia.org/wiki/Anexo:Encuestas_de_intencion_de_voto_para_las_elecciones_presidenciales_de_Argentina_de_2019.

[158] Clarín. *¿Especialistas u operadores? Los consultores quedaron en la mira por el pifie generalizado del 11 de agosto*. 2019. URL: https://www.clarin.com/politica/encuestadoras-fuego-erraron-paso-dicen-octubre_0_T72H9hdl.html.

[159] Clarín. *Intrigas en la Casa Rosada y pases de factura en la City por el “lunes negro”*. 2019. URL: https://www.clarin.com/opinion/intrigas-casa-rosada-pases-factura-city-lunes-negro_0_jnggAIsh5.html.

[160] Impulso Baires. *Bomba: Inversores demandarían a Elypsis por la encuesta del viernes, y los K cargarían con denuncia penal para investigar a operadores*. 2019. URL: https://www.impulsobaires.com.ar/nota/275336/bomba-inversores-demandarian-a-elypsis-por-la-encuesta-del-viernes-y-los-k-cargarian-con-denuncia-penal-para-investigar-a-operadores.

[161] Rachael Levy. *Hedge Fund Loses 1 Billion in One Month on Argentina Bet*. 2019. URL: https://www.wsj.com/articles/hedge-fund-loses-1-billion-in-one-month-on-argentina-bet-11567696547.

[162] Financial Times. *Autonomy Capital lost 16% in Argentina market rout*. 2019. URL: https://www.ft.com/content/29764546-c821-11e9-a1f4-3669401ba76f.

[163] Michael P Battaglia et al. "Tips and tricks for raking survey data (aka sample balancing)". In: *Abt Associates* 1 (2004), pp. 4740–4744.

[164] David Izrael, David C Hoaglin, and Michael P Battaglia. "A SAS macro for balancing a weighted sample". In: *Proceedings of the twenty-fifth annual SAS users group international conference*. Citeseer. 2000, pp. 9–12.

[165] SEIDO. *SEIDO - Special Report: Lie to Me*. 2019. URL: https://us3.campaign-archive.com/?e=&u=e02ede36ce39515be5fb17728&id=3bf5cf2e90.

[166] Ivar Krumpal. "Determinants of social desirability bias in sensitive surveys: a literature review". In: *Quality & Quantity* 47.4 (2013), pp. 2025–2047.

[167] Yuhao Wu et al. "A novel framework for detecting social bots with deep neural networks and active learning". In: *Knowledge-Based Systems* 211 (2021), p. 106525.

[168] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[169] Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018.

[170] Juan Martinez-Romo et al. "Disentangling categorical relationships through a graph of co-occurrences". In: *Physical Review E* 84.4 (2011), p. 046108.

[171] Andrea Ceron, Luigi Curini, and Stefano M Iacus. "Using sentiment analysis to monitor electoral campaigns: Method matters: evidence from the United States and Italy". In: *Social Science Computer Review* 33.1 (2015), pp. 3–20.

[172] Face". *Face",Cognitive Services*. URL: https://www.faceplusplus.com/.

[173] New York Times. *Latest Election Polls 2016*. 2016. URL: http://www.nytimes.com/interactive/2016/us/elections/polls.html.

[174] Jure Leskovec and Christos Faloutsos. "Sampling from Large Graphs". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ?06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, 631?636. ISBN: 1595933395. DOI: 10.11451150402.1150479.

[175] David Gfeller and Paolo De Los Rios. "Spectral Coarse Graining of Complex Networks". In: *Phys. Rev. Lett.* 99 (3 Aug. 2007), p. 038701. DOI: 10.1103/PhysRevLett.99.038701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.038701.

[176] M. Ángeles Serrano, Dmitri Krioukov, and Marián Boguñá. "Self-similarity of complex networks and hidden metric spaces". In: *Physical Review Letters* 100 (7 2008), p. 078701. DOI: 10.1103/PhysRevLett.100.078701.

[177] Mikael Huss and Petter Holme. "Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks". In: *IET Systems Biology* 1.5 (2007), pp. 280–285. DOI: 10.1049/iet-syb:20060077.

[178] Caroline Buckee. "Improving epidemic surveillance and response: big data is dead, long live big data". In: *The Lancet Digital Health* 2.5 (2020), e218–e220.

[179] Raj Kumar Pan and Jari Saramäki. "Path lengths, correlations, and centrality in temporal networks". In: *Physical Review E* 84.1 (2011), p. 016105.