Using Big Data and Network Theory to Inform Decision-making on COVID-19 in Bogotá

A framework for experimental design of policy evaluation of urban interventions

Felipe González-Casabianca[1], Andrea Parra-Salazar[1], Juana Salcedo-Ortiz[1], Federico Andrade-Rivas[7,8], Pablo Cárdenas[3], Alvaro Morales[1], Juliana Maria Damelines-Pareja[9], Diana Sofía Rios-Oliveros[9], Carolina Salazar[1], Santiago Usma[12], Marina Muñoz[4], Luz H. Patiño[4], Nathalia Ballesteros[4], Juan David Ramírez[4], Andrés Ángel[6], Tomás Rodríguez[13], Jaime Cascante[14], Hector Galindo-Silva[11], Stephanie Majerowicz[10], Vladimir Corredor[5], and Alejandro Feged-Rivadeneira[1,*]

[1]DataLama - Facultad de estudios internacionales, políticos y urbanos, Universidad del Rosario, Bogotá D.C., Colombia
[3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Centro de Investigaciones en Microbiología y Biotecnología-UR (CIMBIUR), Facultad de Ciencias Naturales, Universidad del Rosario, Bogotá, Colombia
[5]Departamento de Salud Pública, Universidad Nacional, Bogotá D.C., Colombia
[6]Departamento de Matemáticas, Universidad de los Andes, Bogotá D.C., Colombia
[7]School of Population and Public Health, University of British Columbia, Vancouver, Canada
[8]Instituto de Salud y Ambiente, Universidad El Bosque, Bogotá D.C., Colombia
[9]Subsecretary of Public Health, District Secretary of Health, Bogotá D.C, Colombia
[10]Escuela de Gobierno, Universidad de Los Andes, Bogotá D.C, Colombia
[11]Department of Economics, Pontificia Universidad Javeriana, Bogotá D.C, Colombia
[12]Departmento de Ingeniería Biomédica, Universidad de los Andes, Bogotá D.C, Colombia
[13]Department of Economics, Universidad de los Andes, Bogotá D.C, Colombia
[14]Department of environmental health sciences, mailman school of public health, Columbia university
[*]Corresponding Author, alejandro.feged@urosario.edu.co

ABSTRACT

Supplementary Materials

1 Data Summary

Table 1

Metric		Treatment				Control
		Teusaquillo	Engativá. Vivienda Mixta	Las Ferias	All	All
Number of Devices	Avg	20,977	683	1,385	7,682	785
	Std	18,316	314	457	1,413	498
Number of Contacts	Avg	977,456	18,377	3,674	33,316	2,531
	Std	284,2211	27,693	3,416	169,724	8,308
Average Distance to Infected	Avg	1,917	1,011	1,196	1,375	1,240
	Std	903	545	474	771	726
Personalized PageRank (10^{-5})	Avg	4.76	146.37	72.15	13.01	127.45
	Std	15.87	196.71	113.58	54.33	230.03

2 Personalized PageRank and multiSIR model

We study a Multi-Sir model. We focus on approximations to the final size using an implicit first order approximation, Newton-Raphson iterations. We obtain centrality measures calibrated to the Multi-Sir model that are cases of PageRank and Personalized PageRank, we interpret this approximation using random walks.

2.1 Multi-Sir

For a weighted graph we have an edge between node j and i if $a_{ij} > 0$ (note the reverse indices, this is done to be compatible with matrix multiplication), we consider a Multi-Sir deterministic model using the adjacency matrix a_{ij}

$$S_i' = -\beta S_i \sum_{j=1}^{n} a_{ij} I_j$$

$$I_i' = -\gamma I_i + \beta S_i \sum_{j=1}^{n} a_{ij} I_j$$

$$R_i' = \gamma I_i$$

2.1.1 Final Size

We focus on the final size on each node:

$$\lim_{t\to\infty} R_i(t)$$

Consider the function

$$\iota_i(t) = \sum_{j=1}^{n} a_{ij} R_j(t)$$

and

$$\frac{dS_i}{d\iota_i} = \frac{S_i'}{\iota_i'} = \frac{-\beta S_i \sum_{j=1}^{n} a_{ij} I_j}{\gamma \sum_{j=1}^{n} a_{ij} I_j} = -\frac{\beta}{\gamma} S_i = -R_0 S_i$$

giving

$$\ln(S_i(t)) = -R_0 \iota_i(t) + K$$

using the initial condition $R_i(0) = 0$.

$$\boxed{S_i(t) = S_i(0) e^{-R_0 \iota_i(t)} = S_i(0) e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(t)}}$$

We have $S_i + I_i + R_i$ is constant, and when $t \to \infty$, $I_i(t) \to 0$ and $\lim_{t\to\infty} R_i(t)$ and $\lim_{t\to\infty} S_i(t)$ exist, therefore

$$S_i(\infty) = S_i(0) e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)}$$

We call

$$N_i = S_i + I_i + R_i$$

giving the final size equation

$$\boxed{R_i(\infty) = N_i - S_i(0) e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)}}$$

The total final size of the multi-sir model is

$$R = \sum_i R_i(\infty)$$

From now on, we will interpret (S_i, I_i, R_i) as probabilities and therefore $N_i = 1$.

2.2 First derivatives of the final size and first order approximation

Now we study the approximation of the final size with first order approximation. The partial derivatives with respect to $S_k(0)$

$$\frac{\partial R_i(\infty)}{\partial S_k(0)} = -\delta_{ik} e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)} - S_i(0) e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)} \left(-R_0 \sum_{j=1}^{n} a_{ij} \frac{\partial R_j(\infty)}{\partial S_k(0)} \right)$$

which can be written (when $S_i(0) \neq 0$)

$$\frac{\partial R_i(\infty)}{\partial S_k(0)} = -\delta_{ik} \frac{S_i(\infty)}{S_i(0)} + S_i(\infty) R_0 \sum_{j=1}^{n} a_{ij} \frac{\partial R_j(\infty)}{\partial S_k(0)}$$

Lets call $r_{ik} = \frac{\partial R_i(\infty)}{\partial S_k(0)}$ then we have a system

$$r_{ik} = -\delta_{ik} \frac{S_i(\infty)}{S_i(0)} + R_0 \sum_{j=1}^{n} \tilde{a}_{ij} r_{jk}$$

where $\tilde{A}$ is the matrix $Diag(S(\infty))A$, which can also be written in matrix form as:

$$(I - R_0 \tilde{A}) r = -Diag(\frac{S_i(\infty)}{S_i(0)})$$

This, we have obtained that the derivatives of the final size with respect to initial conditions $S_k(0)$ is:

$$\boxed{r = -(I - R_0 \tilde{A})^{-1} Diag(\frac{S_i(\infty)}{S_i(0)})}$$

where $\tilde{A} = Diag(S(\infty))A$.

Lets suppose all initial probabilities of being infected are the same, that is: $I_i(0) = \alpha$, then $S_i(0) = 1 - \alpha$ and we can write a vector of derivatives

$$\boxed{\frac{dR_i(\infty)}{d\alpha} = -\frac{1}{\alpha}(I - R_0 \tilde{A})^{-1} \vec{S}(\infty)}$$

Which shows the sensitivity of the nodes to a change in the initial susceptible probabilities. (The negative sign comes from the fact that we are derivating against initial susceptibles = 1-initial infected).

The problem with this formula is that still needs $S_i(\infty)$, we can use this derivative near the disease-free equilibrium (DFE) to find a more useful approximation.

2.2.1 Disease-Free Equilibrium

When $I_i(0) = 0$, then $S_i(0) = S_i(\infty) = 1$, we can approximate near this solution. At DFE

$$\tilde{A} = A$$

The total final size can be approximated near the DFE using $R_i(\infty) \approx 0 + \sum_k \frac{\partial R_i(\infty)}{\partial S_k(0)} \Delta S_k(0)$

$$R = \sum_i R_i(\infty) \approx 0 - \vec{1}^T (I - R_0 A)^{-1} (\Delta \vec{S}(0)) = -\Delta \vec{S(0)}^T (I - R_0 A^T)^{-1} \vec{1} = \Delta \vec{I(0)}^T (I - R_0 A^T)^{-1} \vec{1}$$

which is probably better written in terms of fractions of initial susceptible,

Lets set all $\Delta S_k(0) = \alpha$ and get

$$R \approx \alpha \sum_k \left(\sum_i \frac{\partial R_i(\infty)}{\partial S_k(0)} \right)$$

and we can ask which node is contributing more to this quantity, i.e. find k such that

$$\sum_i \frac{\partial R_i(\infty)}{\partial S_k(0)}$$

is maximum.

$$\sum_i \frac{\partial R_i(\infty)}{\partial S_k(0)} = \vec{1}^T (I - R_0 A)^{-1} e_k = e_k^T (I - R_0 A^T)^{-1} \vec{1}$$

is the PageRank centrality of the k-th node of the graph with adjacency matrix A^T with parameter R_0.

2.2.2 Newton Raphson iterations

In the previous section we used a simple first order approximation, but since the final size can also be written as the zero of a vector valued function we can use the Newton-Raphson iterations to approximate them.

We rewrite

$$\boxed{R_i(\infty) = 1 - S_i(0)e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)}}$$

as

$$\boxed{R_i(\infty) + S_i(0)e^{-R_0 \sum_{j=1}^{n} a_{ij} R_j(\infty)} - 1 = 0}$$

recall that Newton-Raphson is an iterative method for approximating $f(\vec{x}) = \vec{0}$, we start with a well-choosen $\vec{x}_0$ and then approximate $f(\vec{x}) \approx f(\vec{x_n}) + Df(\vec{x_n}) \cdot (\vec{x} - \vec{x_n})$ to obtain

$$\vec{x_{n+1}} = \vec{x_n} - Df(\vec{x_n})^{-1} f(\vec{x_n})$$

this is quadratic method with guaranteed convergence for a good choice of $\vec{x_0}$.

In our case $f_i(\vec{x}) = x_i + S_i(0)e^{-R_0 \sum_{k=1}^{n} a_{ik} x_k} - 1$, the derivative is

$$\frac{\partial f_i}{\partial x_j} = \delta_{ij} - S_i(0) R_0 e^{-R_0 \sum_{k=1}^{n} a_{ik} x_k} \sum_{k=1}^{n} a_{ik} \delta_{kj} = \delta_{ij} - S_i(0) R_0 e^{-R_0 \sum_{j=1}^{n} a_{ij} x_j} a_{ij}$$

Lets look at the initial iteration taking $\vec{x}_0 = \vec{0}$, for which $f(\vec{x_0}) = \vec{1}$,

$$Df(\vec{x_0}) = I - R_0 A$$

which recovers the approximation from the previous section.

In general, lets suppose we take $\vec{x_n} \approx \vec{R}(t_n)$ for some times t_n, then

$$Df(\vec{x_n}) = I - R_0 \tilde{A}$$

where $\tilde{A} = Diag(S(t_n))A$, and

$$f(\vec{x_n}) = R(t_n) + S(t_n) - 1 = -I(t_n)$$

therefore the iterations for Newton-Raphson are approximately

$$R(t_{n+1}) \approx R(t_n) + (I - R_0 \tilde{A})^{-1} (I(t_n))$$

we can reformulate this as giving the change of probability to become recovered between t_n and t_{n+1}:

$$\Delta R \approx (I - R_0 \tilde{A})^{-1} (I)$$

with $\tilde{A} = Diag(S(t_n))A$, which can be thought as a Personalized PageRank, where the personalization is given by the probabilities of being infected.

Suppose for a moment that at t_n all the nodes have the same probabilities: to be susceptible α and probability β to have recovered and therefore $1 - \alpha - \beta$ to be infected, then the probability to become recovered between t_n and t_{n+1} is approximately given by:

$$(I - \alpha R_0 A)^{-1} (\overline{1 - \alpha - \beta}) = (1 - \alpha - \beta)(I - \alpha R_0 A)^{-1} (\bar{1})$$

which is a multiple of PageRank of the matrix A with parameter αR_0.

2.2.3 Integrating incomplete information

When we have additional information about the infected ones we can use our previous approximation. The probability to become recovered between t_n and t_{n+1} is given by:

$$\Delta R \approx (I - R_0 \tilde{A})^{-1} (I(t_n))$$

with $\tilde{A} = Diag(S(t_n))A$.

Now suppose that a given time, somehow we acquire an estimate of the probability of each node to be infected but we do not know the probabilities of being infected before, therefore for simplicity (or lack of better knowledge) we are going to assume that the probabilities of being susceptible are all the same $S(t_n) = \alpha$, a number small enough. Therefore, the change on probability to become recovered between t_n and t_{n+1} is

$$(I - \alpha R_0 A)^{-1}(I(t_n))$$

which is a multiple of Personalized PageRank.

We can state our intuitive interpretation of our approximation:

Given some good estimate on the probability of each node to be infected and assuming equal small enough probability to be susceptible for all nodes, Personalized PageRank gives a multiple of the probability of each node to be infected at some time due to network effects.

Some more explanation:

- Good estimates implies that we are in a region where Newton-Raphson is converging quadratically, this implies that the next iteration corresponds to a t_{n+1} very big compared to t_n, so we can ignore the time an infected person needs to recover and thus $R(t_{n+1}) - R(t_n)$ is approximately the probability to be infected between t_n and t_{n+1}, which basically means the probability to be infected at some point after t_n.

- We choose α, the probability of being Susceptible small enough, so that for all nodes

$$1 \geq 1 - \alpha - I_i(t_n) \geq 0$$

which leads to assume that

$$1 - \max_i I_i(t_n) > \alpha > 0$$

which is needed to make sense of $R_i(t_n) = 1 - \alpha - I_i(t_n)$ as a probability to be Recovered.

2.2.4 Spatial information

Suppose that we have partial (and imperfect) spatial information on the infected ones. We are not going to assume that we know exactly which nodes are infected, but are going to assign a probability to be infected dependent on the spatial distance to the infected ones.

We define the probability of an individual to be infected

$$I_i(t) = \frac{\ln(1 + e^{-d_i})}{\gamma}$$

where d_i is the spatial distance from node i to the closest infected case and γ is a parameter, larger or equal to $\ln(2)$. $\gamma = \ln(2)$ was chosen so the nodes at distance zero from infected cases have probability one of being infected

3 Simulations

As noted in Section 2, the vector of Personalized PagerRanks of the agents is a multiple of the vector of probabilities that the agents become infected at some point. In order to gauge the quantitative implications of the main findings reported in Table 1 of Section 6.1, we ran 100 simulations using a computational SIR model involving 1000 agents who live and move stochastically on and a grid, and which come in contact with each other according to a network generated based on their proximity on the grid, over 80 time periods. For each time period in each simulation, we computed the mean probability of infection of susceptible agents during the following 7 time periods (*the probability of next-week infection*), conditional on their Personalized Page Ranks. Note that while the total PageRank is by construction always equal to 1, and thus its mean in the population is $1/1000$, its distribution varies widely as the virus spreads throughout the population. It follows, that its variance, as well as the impact on the probability of next-week-infection changes throughout the simulated epidemic. Figure **??** shows the mean probability of infection over all simulations of susceptible nodes within seven days (periods), starting on the day shown in the axis. The red line shows this mean probability for susceptible agents whose PPRs were within 0.1 standard deviations of the mean. The blue line shows this mean probability for susceptible agents whose PPRs were between -0.1 and -0.2 standard deviation from the mean. Note that the difference between the mean probabilities is between 1 and 2 percentage points and varies widely throughout the epidemic.

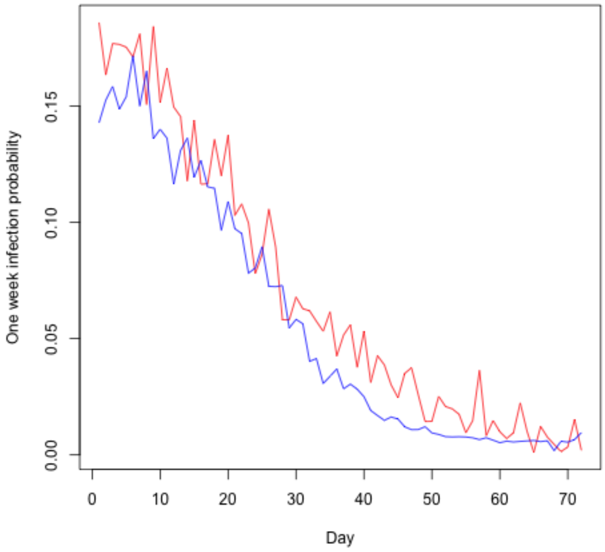

Figure 1. Mean probability of infection of susceptible nodes within seven days (periods), after the day shown in the axis. The red line shows the mean probability of infection among susceptible agents whose PPRs were ±0.1 standard deviations of the mean (1/1000). The blue line shows the mean probability of infection of susceptible agents whose PPRs were between -0.1 and -0.2 standard deviation from the mean.

4 Genetic Sampling

A total of 377 SARS-CoV-2 genome sequences of Colombian origin were sequenced through this project, obtained from samples provided by Grupo de Investigación en Enfermedades Tropicales del Ejército (GINETEJ), Laboratorio de Referencia e Investigación, Dirección de Sanidad Ejército. Sample collection was carried out by nasopharyngeal swabs of individuals sampled from the Military Hospital in Bogotá. Individuals provided informed consent and sample collection was approved by Universidad del Rosario's Research Ethics committee (DVO005 1550-CV1400).

Whole genome sequencing of SARS-CoV-2 was performed using Oxford Nanopore's MinION platform, using the MinKNOW application (v1.5.5) according to the established protocol (https://artic.network/ncov-2019). The bioinformatic analysis was performed on the raw Fast5 files, which were basecalled to obtain the Fastq files, and then demultiplexed using the Guppy tool. After that, the Fastq files already assigned by barcode were filtered by quality and length eliminating possible chimeric reads. Finally, genome assemblies were obtained following the algorithm for MinION sequences described in the ARTIC bioinformatics pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). Once the assemblies were obtained, typing was performed based on the Pangolin COVID-19 Lineage Assigner (Phylogenetic Assignment of Named Global Outbreak LINeages). The mutation search was performed by means of Clade assignment, mutation calling, and sequence quality checks NextClade v 1.5.4 (https://clades.nextstrain.org/).This yielded 377 SARS-CoV-2 genome sequences. Raw sequences are deposited on GISAID[1] (gisaid.org).

Genomes were preprocessed, aligned, and included in a maximum-likelihood phylogeny of subsampled global SARS-CoV-2 genomes using Nextclade (clades.nextstrain.org) under default parameters, part of the Nextstrain-TreeTime computational framework[2,3]. The resulting phylogeny was visualized using Auspice (auspice.us)[2] with custom colors and region implemented independently.

Two genomes sequenced in this study (samples A339 and A514) were found to cluster poorly with other samples and were flagged by Nextclade as having poor quality private mutations. The genomes were found to combine sequence similarity to different lineages, including lineages not in existence at the time of collection (sample A339 was collected in 11 April 2021, almost six months prior to the emergence of Omicron REF, with which it holds significant similarity). These sequences were deemed likely to be the result of cross-contamination events between samples, and not considered in further analysis.

The remaining 375 genomes were plotted on a phylogeny and represent sequences from variants of concern Alpha, Gamma, Lambda, Delta, Mu, and Omicron throughout the course of the pandemic. Genetic distance was calculated as the Kimura

2-parameter distance, based on the APE R package implementation of molecular models of evolution for genetic distance[4].

5 Device co-location as a proxy for contacts

One of the key assumptions on co-location or digital contact based methodologies is that the device interaction is indeed a good proxy for physical contacts between users. Although it has been shown that these techniques have their limitations[5], specially when it comes to the selection of the interaction distance parameter[6], a link between between digital proximity and effective public policy has also been established[7].

One way to justify the use of device co-location as an effective proxy for actual contact, at least in the context this study, was to evaluate the correlation of the distance between the devices of two COVID-19 positive users, with the genetic distance from their sequenced samples. Specifically, given a couple of geolocated COVID-19 samples g_i and g_j taken from different individuals, we wanted to see if there is correlation between:

- **Genetic distance:** the Kimura 2-parameter distance as mentioned in 4.
- **Device distance:** using the geolocation of sample g_i, we obtained the set D_i of devices that most likely correspond to the subject's actual device. The set $D_i = d_1, d_2, \ldots d_n$ contains all the devices who's mobility patterns show that they had spend at least one night 30 meters or less from the device, in the previous month since the subject's first symptoms appear. Also, a network of devices was built between all devices that where active in the previous 14 days of the sample's first symptoms date across the Bogotá's metropolitan area. The nodes and edges of the network where computed using the same schema as the main methodology (see section main manuscript). So the distance $\delta(d_k, d_l)$ between devices d_k and d_l corresponds to the length of the shortest path between them over the constructed network with all edges having an equal weight of 1. If d_k and d_l are in different connected components of the graph, then $\delta(d_k, d_l) = \infty$. Thus, the device distance between g_i and g_j is:
$$\Delta(g_i, g_j) = \min_{d_k \in D_i, d_l \in D_j} \delta(d_k, d_l)$$

Although theoretically possible to correlate these two distances, we encountered a sampling problem that we wish to discuss. As mentioned in 4, we started with 377 samples, of which only 168 (44.5 %) had sufficient metadata to be geolocated inside the metropolitan limits. From those remaining samples, we were only able to associate 34 (9.02 %) to devices and from those corresponding devices, only 3 devices (0.7 %) were inside a connected component graph. This means that only three samples have non infinite device distance, which renders useless the correlation study with the obtained genetic distances.

This serves to show the importance of proper articulation with the local government, more so when genetic sampling is required. It is crucial to have in place a robust data and sample recollection scheme, where informed consents are always included (we had to discard significant amount of samples because the informed consents were missing or were not handed) and metadata quality and availability are regarded as priorities.

Another negative factor that could be improved is sequencing a useful subset of samples. Since we could only sequence a fraction of the collected samples, it is crucial to select the most relevant samples, which in our case means samples that can be geolocated, associated to a device and that the device has an important role (i.e high centrality) in a connected part of the network. Since most of the available samples for sequencing were either discarded or lacked the consent form, our final samples represented a very specific time and place in the pandemic, making the proposed analysis even more difficult. With correct sample recollection protocols, where all patients were offered a consent form and addresses were filled in correctly, we could select the key samples to be sequenced and compare the two defined distances to further confirm our methodology and choice of metrics.

References

1. Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

2. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

3. Sagulenko, P., Puller, V. & Neher, R. A. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution* **4**, vex042 (2018).

4. Paradis, E., Claude, J. & Strimmer, K. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290 (2004).

5. Génois, M. & Barrat, A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.* **7**, 1–18 (2018).

6. Stopczynski, A., Pentland, A. S. & Lehmann, S. Physical proximity and spreading in dynamic social networks. *arXiv preprint arXiv:1509.06530* (2015).

7. Cencetti, G. *et al.* Digital proximity tracing on empirical contact networks for pandemic control. *Nat. communications* **12**, 1–12 (2021).