PROPOSAL FOR COVID-19 MODELING IN MALAYSIA

DAVID BENJAMIN LIM

Abstract. We suggest three situations to be modeled in order to understand COVID-19 in Malaysia. We implement the first of these and show that even though workplace clusters contributed to only 15% of all COVID-19 cases from 1/6/2021 to 23/7/2021, their contribution to the total R_0 of Malaysia was so significant that *if* there were no such clusters, the total R_0 of Malaysia would have been 0.9. This is strong evidence to suggest that workplace clusters were the driving force behind the increase in COVID-19 in Malaysia during this time period.

Contents

1. Introduction

The COVID-19 pandemic has forced governments around the world to take measures to contain the spread of the disease. In Malaysia, the government has already implemented several rounds of lockdowns, the most recent of which is the FMCO (full movement control order) which started on 1/6/2021. The FMCO is to this date mostly at full strength albeit for a few states that have been scheduled to move on to what the Malaysian government calls "Phase 3" (Sarawak, Perlis and Labuan). During this phase, things such as indoor dining and gyms are allowed to operate among other things.

For any action taken to control the spread of COVID-19, there is an associated cost. Businesses ordered to close result in lost income and in worst-case scenarios the destruction of careers that have been built over many years, while sedentary lifestyles associated with stay-at-home orders impact mental/physical health. In addition, because any policy taken to control the spread of COVID-19 is only effective if compliance is high, any policy decided by the government must necessarily factor human behavior into the equation.

In view of the paper [CCI+21], which finds that the targeted policies suggested in *loc. cit.* Pareto dominate a blanket lockdown strategy, we suggest that the government of Malaysia adopt a targeted approach in dealing with the COVID-19 pandemic in Malaysia. It is clear that in order to do this, it is necessary to know the *driving force* behind the pandemic in Malaysia (see Section 3 and Appendix A). We calculate for Malaysia during the time period 1/6/2021 to 23/7/2023 that the contribution to R_0 during this time period from workplace clusters is significant enough to obtain strong evidence that *if* there were no workplace clusters, the R_0 of Malaysia during this time would have been less than 1 (see Section 3 for details). In other words, COVID-19 would have *decreased*.

Date: August 6, 2021.

Along with such a targeted strategy, we urge that it is crucial to be able to estimate the number of ICU beds needed at any given time. Such an estimate would allow the government to plan ahead of time in the event more ICU beds are needed, avoiding knee-jerk reactions which themselves have associated costs. Our suggestion to estimate the number of ICU beds needed is outlined in Section 4. Finally, in order to be able to contain COVID-19 in Malaysia, it is necessary to know the "true" extent of COVID-19 in Malaysia. Our suggestion to estimate this is the content of Section 5 and Appendix B.

2. Acknowledgments

We thank Eric Cooper of the Schleier-Smith Lab, Stanford University and Samuel Tenka of the Computer Science and Artificial Intelligence Lab, MIT for many insightful discussions concerning the mathematics behind COVID-19. We also thank Yang Liu (Stanford University), Mark Sellke (Stanford University) and Brian Lawrence (University of Chicago) for many helpful discussions. We are also grateful to have received the support of Latifah Hani Hamzah (Stanford University).

3. What is driving the spread of COVID-19 in Malaysia?

In a speech in parliament on 27/7/2021, the minister of international trade and industry, Dato' Seri Azmin Ali said that the manufacturing sector contributed to only 8.8% of all cases between 1/6/2021 and 23/7/2021.[1]After saying this, he says (in Malay): "Saya boleh bagi lagi sampling, Datuk Speaker sama ada harian, mingguan atuapun bulanan, yang membuktikan sumbangan sektor perkilangan kepada kes-kes harian (mic cuts out)."[2] In english, this translates to: "I can give you a sampling, Datuk Speaker be it daily, weekly or monthly, that proves that the contribution of the manufacturing sector to daily cases (mic cuts out)." Although the rest of this sentence cannot be heard, it can be inferred that the minister claims that because the manufacturing sector only contributed to 8.8% of all COVID-19 cases (during the time period mentioned above), that it *cannot* be the cause of the spread of COVID-19 in Malaysia.

However, this line of reasoning is wrong: More precisely, just because a particular group of people make up a small percentage of all COVID-19 cases, *does not mean* that they *cannot* be the source of the problem (see Appendix A). The reason for this is that it is entirely possible for there to be a small subset of the population that is very infections (i.e. have very high R_0), and therefore contributes significantly to the *total* R_0 of Malaysia (Theorem A.1). To be maximally precise, we do not claim that the manufacturing sector *is* the source of the COVID-19 problem in Malaysia. All we are doing is simply pointing out that the line of reasoning employed by Dato' Seri Azmin Ali in his speech in parliament is incorrect. Therefore, we suggest a thorough and systematic investigation into the driving force behind the spread of COVID-19 in Malaysia.

Let us show using the ideas in Appendix A that the total R_0 from workplaces throughout Malaysia from 1/6/2021 to 23/7/2021 is big enough that if we *substract* it from the R_0 for the whole of Malaysia during this time period, we get a number less than 1, i.e. COVID-19 would have died decreased. Indeed, the total R_0 of Malaysia during this time period was 1.0764, while that of workplaces was 0.1712. Therefore, their difference is ≈ 0.9 which is a number less than 1 that is significant enough to give strong evidence to the following: If there were no workplace clusters during this time period, COVID-19 would have decreased, *even though* workplace clusters only accounted for approximately 15% of all cases. We refer the reader to Section A.3 for details.

4. Estimating the number of COVID-19 ICU beds needed

Given the gravity of the COVID-19 situation in Malaysia, it is clear that there is a need to model the expected number of ICU beds needed. Currently, to the best of our knowledge, no such thing is done, at least for the state of Penang.[3] We would like to propose a simple model to be able to estimate, given the current number of ICU cases and breakdown of daily new cases on a given day, the number of ICU beds needed say 7 or 14 days from now.

[1]Around the 3:07:00 mark here

[2]Around the 3:07:10 mark here

[3]This was confirmed by the Chief Minister of Penang during a zoom meeting with YB Sim Tze Tzin and myself on 22/7/2021.

This model is probabilistic in nature in the sense that it is based on the idea that for a given person X with COVID-19, there is an associated probability P_X of ending up in ICU. This probability P_X is a function of X's age, vaccination status, comorbidities etc and can be estimated to first-order using simple statistical techniques. We stress that some amount of feature selection must be done in order for such a model to be meaningful, because certain features are correlated, e.g. being older and having a comorbidity. Furthermore, the selection of "correct" features to be used in such a model necessarily requires domain knowledge about the nature of the SARS-CoV-2 virus. For a simple example to illustrate this, it is known empirically that the probability that a person with COVID-19 will end up in ICU is very heavily dependent on age (in fact the probability distribution is highly skewed to those over 60). This alone suggests that the feature of daily new cases be at least further sub-divided into those above 60 and under 60.

5. Estimating the true number of cases

This last section is somewhat theoretical in nature. However, it is motivated by the following question which we believe to be of tremendous interest to health experts and epidemiologists in Malaysia.

Question: Given the number of tests on a given day in Malaysia, as well as the number of new detected COVID-19 cases, what is the "true" number of COVID-19 cases?

We "know" and might expect that if the test positivity rate is more than 10% that the true number of cases is much higher than those detected. But by how much? If tests are random (in probability theory terms i.i.d. or independent and identically distributed) then it is possible to bound the probability that the difference between the true and test positive rate differs by more than $\varepsilon = 0.01$ say, as a function of the number of tests and test positive rate (Appendix B). However, in real life tests are almost never random simply because if one person has COVID-19, all of that person's close contacts are tested as well.

Let $X_1, \ldots, X_M$ denote the COVID-19 tests done on a given day. These are not i.i.d. but there may be hidden latent variables $Z_1, \ldots, Z_K$ that "explain" the X_i's. For instance, the case of a single latent variable Z (more precisely the *event* $Z = 1$) may encode whether or not there is an outbreak at a factory. What I expect to be true is a similar concentration type bound for hidden Markov models like this as in the case of i.i.d. tests. I have written down the i.i.d. case but have not worked out the general case of hidden latent variables. Again, we stress that it would be very helpful to have real input from those with specialties in modeling infectious diseases, in order for the assumptions of such a model to be somewhat justified with real-world reasons.

6. Conclusion

In summary, we suggest three things that the government of Malaysia must do:

(1) Undertake a thorough and systematic investigation into the driving factors behind the spread of COVID-19 in Malaysia.
(2) Estimate the number of COVID-19 ICU beds needed in the future.
(3) Estimate the true number of COVID-19 cases in Malaysia.

The contents of this paper lay out a plan to execute these ideas. Note however that the quality of any model produced depends on the quality of data available. The current data available on the Malaysian Ministry of Health GitHub is not fine enough to be able to implement an accurate model for predicting ICU beds along the lines of Section 4. Therefore, we welcome and urge the MOH to reveal data at a finer level than what is currently available to the public.

Appendix A. Measuring the total contribution of a particular sector to the total R_0

A.1. **Theoretical Foundations.**

Theorem A.1. Consider the random tree generated by the following algorithm:

```
# Initialize k nodes
# Initialize number of generations n
# Initialize probabilities p and q with p + q = 1
# Initialize growth rates R_p and R_q

do n times:
```

```
for node in nodes:
    node.color = red w/ probability p, blue w/ probability q
    if node.color is red:
        add R_p children to node
    else:
        add Pois(R_q) children to node
nodes = nodes.children
```

Then as $n \to \infty$, the expected number of nodes is infinite if $pR_p + qR_q > 1$, and $k/(1-(pR_p+qR_q))$ (in particular finite!) if $pR_p + qR_q < 1$.

Proof. We use 0-indexing. Let X_i be the random variable that counts the number of nodes at generation i; by definition $X_0 = k$. Conditioning on whether a node at generation $i-1$ is blue or red, we have

$$\mathbf{E}[X_i|X_{i-1}] = X_{i-1}(pR_p + qR_q).$$

Hence, by the law of total expectation,

$$\mathbf{E}[X_i] = \mathbf{E}\left[\mathbf{E}[X_i|X_{i-1}]\right] = \mathbf{E}[X_{i-1}(pR_p+qR_q)] = \mathbf{E}[X_{i-1}](pR_p+qR_q).$$

It follows that $\mathbf{E}[X_i] = k(pR_p+qR_q)^i$ and hence

$$\begin{aligned} \text{Expected total nodes after } n\text{-generations} &= \sum_{i=0}^{n-1} \mathbf{E}[X_i] \\ &= \sum_{i=0}^{n-1} k(pR_p+qR_q)^i \\ &= k\left[\frac{(pR_p+qR_q)^n - 1}{(pR_p+qR_q)-1}\right]. \end{aligned}$$

Letting $n \to \infty$, this computation shows that if $pR_p + qR_q > 1$, then on average the tree has an infinite number of nodes, and if it is less than 1, it has $k/(1-(pR_p+qR_q))$ nodes. □

A.2. **Real-world applications of Theorem A.1.** Theorem A.1 above shows the following. It is possible to have a proportion p of red nodes, with $p << q$, but with large R_p in a way that $pR_p + qR_q > 1$. The random tree we generate in Theorem A.1 above will grow indefinitely even though the total proportion of red nodes is small. Furthermore, one may also arrange so that $R_q < 1$, so that if there are no red nodes present (set $p = 0$ and $q = 1$), then the number of nodes in the tree will be finite. This shows that even though the red notes make up a very small portion of the tree, it is precisely these nodes that contribute to the tree growing indefinitely.

Said differently, if a red node represents a workplace cluster and blue node a sporadic case, then it is entirely possible for the number of workplace clusters (red nodes) to be very small, but that it is the main driver behind the spread of COVID-19 in Malaysia. In practice, Theorem A.1 can be used to as follows. Fix some time period (for instance the one mentioned by Dato' Seri Azmin Ali in parliament, i.e. 1/6/2021 - 23/7/2021). Then (for example) red nodes may be workplace clusters and blue sporadic cases. The probability p (and hence $q = 1-p$) can be empirically calculated from the file `epidemic/clusters.csv` in the (public!) GitHub repository of the Malaysian Ministry of Health here. The same is true of R_p, and then $R_q = (R_0 - pR_p)/q$ is calculated using R_p and the total R_0 of Malaysia during this time period. If we find that $R_q < 1$, this gives strong evidence to the fact that if there were *no* workplace clusters during our fixed time period, that COVID-19 would have "died off" in Malaysia, i.e. workplaces were behind the spread of COVID-19.

As a final note, for the purpose of quick and fast empirical estimates, R_0 during this time period may be calculated as follows. Let t_{start} and t_{end} be the start and end of the time period respectively, and $f(t)$ the number of daily new cases reported at time t (The value $f(t)$ can be obtained empirically from the file `epidemic/cases_malaysia.csv` of *loc. cit.*). Then

$$R_0 \approx \exp\left(\frac{\log f(t_{\text{end}}) - \log f(t_{\text{start}})}{\text{length (in days) of time period}} \times \tau\right), \tag{1}$$

where τ is the serial interval of COVID-19.

A.3. **An example calculation.** Consider the time period 1/6/2021 to 23/7/2021. Using the ideas of this appendix, let us show that the contribution of workplaces to the R_0 of Malaysia during this time period is indeed significant enough that upon *substracting* it, we get a number less than 1. We will think of a workplace cluster as corresponding to a single red node, and every other COVID-19 case as a single blue node.

Let us first calculate the probability p as well as the quantity R_p. The probability p, i.e. the probability of a workplace cluster, is simply

$$\begin{aligned} p &= \text{probability of workplace cluster} \\ &= \frac{\#\text{people that started a workplace cluster}}{\#\text{people who could have started a workplace cluster}} \\ &\approx \frac{\#\text{workplace clusters}}{\text{Total cases - Total workplace cases} + \#\text{workplace clusters}}. \end{aligned}$$

On the other hand R_p, the number of children of red node, i.e. the number of cases generated out of workplace clusters, is simply

$$R_p = \frac{\text{Total workplace cases}}{\#\text{workplace clusters}}.$$

Let us now calculate these numbers. We will use the data provided by Dato' Seri Azmin Ali in parliament, which he has been kind enough to upload on his facebook page here. We have:

$$\begin{aligned} \text{Total cases} &= 408,134 \\ \text{Total workplace cases} &= 60,587 \\ \#\text{workplace clusters} &= 743 \end{aligned}$$

from which it follows that $p = 0.0021$, $R_p = 81.5437$. Hence the *total contribution* to the R_0 of Malaysia from workplaces is

$$pR_p = 0.1712.$$

On the other hand, by the method of least squares, we find that the R_0 for Malaysia is

$$R_0 = 1.0764.$$

The reader may obtain a similar value without the method of least squares using the estimate (1). Hence

$$R_q = \frac{R_0 - pR_p}{1-p} = 0.907.$$

In summary, even though workplace cases only accounted for $60,587/408,134 \approx 15\%$ of all cases, their contribution to the R_0 is sufficiently significant to make the argument that without workplace clusters, COVID-19 would have decreased in Malaysia.

Appendix B. Measuring the difference between the true and test positive rate

Suppose we have a population of N people and we conduct M i.i.d. tests with replacement. Let X be the discrete random variable that counts the number of positive tests, and Y the continuous random variable that counts the *proportion* of true COVID-19 cases in the population (so that Y is valued in $[0,1]$). For a given number of positive tests m, note that the conditional random variable $Z := (Y|X = m)$ is exactly the true fraction of positive cases, *given that* we have m positive tests. Therefore, the question of how far away the true positive rate is from the positive rate, reduces to the following question:

Question B.1. Given a threshold $\varepsilon > 0$, how can we bound the probability $\Pr(|Z - m/M| > \varepsilon)$?

We will use standard techniques from Bayesian statistics to answer this question. As usual, we will need a prior on Y, which for now we just take to be uniform. In this situation, a straightforward calculation with Bayes' Theorem shows that

$$Z \sim \text{Beta}(m+1, M-m+1).$$

Note that $\mu := \mathbf{E}[Z] = (m+1)/(M+2)$ which is very close to the test positivity rate m/M. Therefore, the question above reduces to bounding the probability $\Pr(|Z - \mu| > \varepsilon)$. Using the fact that the Beta

distribution admits a description in terms of order statistics of uniform random variables, and applying the multiplicative form of Chernoff's inequality, we obtain the concentration bound

$$\Pr(|Z-\mu|>\varepsilon) \leq 2\exp\left(-\frac{(M+1)\varepsilon^2}{3(\mu+\varepsilon)}\right). \tag{2}$$

Note very crucially that the bound on the right does not depend on the size of the population. For concreteness, if $\varepsilon = 0.01$, we test $M = 10000$ people and obtain 1000 positive tests, then the right side of (2) evaluates to approximately 0.097, i.e. there is only a 9.7% chance that the true positive rate differs from the test positive rate by *more* than 1%.

References

[CCI+21] Sergio Camelo, Dragos F. Ciocan, Dan A. Iancu, Xavier S. Warnes, and Spyros I. Zoumpoulis, *Quantifying the benefits of targeting for pandemic response*, medRxiv (2021).