Data-based investigation of COVID-19 topic modelling and mobility patterns in Ontario, Canada

Colin Pierce
York University
Toronto, Ontario
cbpierce@yorku.ca

ABSTRACT

Google's Community Mobility Reports (GCMR) have, in the past, been used to analyze COVID-19 case incidence with respect to changes in mobility. We adopt this technique with the caveat of including analysis for mutant strains and make predictions beginning after the third major closure in Ontario for all public health units. The results are interpreted for commonalities and differences and demonstrate a stronger correlation (positive and negative) between grocery and residential type cases.

A new approach to topic modeling is also outlined based on the web-crawling of Wikipedia article titles. The results are graphed and pruned to discern underlying structures. It is found that *Social Distancing* and the *World Health Organization* are well-connected.

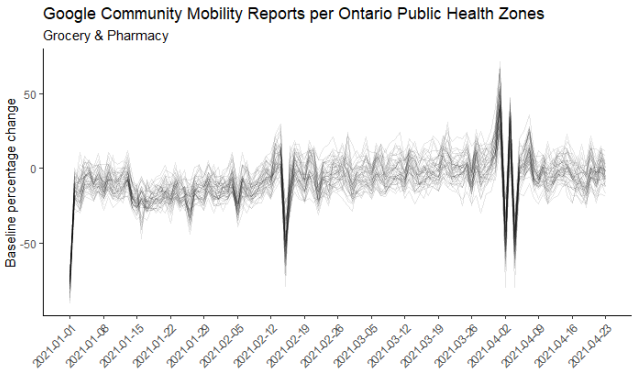

Figure 1: Grocery and pharmacy mobility changes since the new year. The periodic troughs should be interpreted as the 'weekend effect.'

1 INTRODUCTION

Since March 2020, Google has released and maintained a database of mobility activity for each country called the *Community Mobility Reports* (GCMR). It includes, by day, the baseline percentage change in the frequency of six activity categories: Retail, Grocery, Parks, Transit, Work, and Residential; the baseline was calculated between the 5-week period in 2020, January 3rd to February 6th. Initial analysis yielded expected correlations between the differing categories and case rates. [4] However, developments have accrued since the beginning of the pandemic such as the spread of mutant strains and civil unrest that make it worth revisiting.

Moreover, the Ford government has been repeatedly criticized for its binary, province-wide approach to lockdowns. By analyzing the correlation for mobility by case rate per public health unit, we may discern which places need it more than others and hopefully improve insight into decision making.

For more general insight into the pandemic, people often turn to the internet for their learnings. Given the large volume of consumption, then, it poses the question as to whether there exist any biases in our representation of the facts, i.e., do there exist instances of the over-reporting of certain subjects? Given the modern state of the internet, Wikipedia may be the best representation of the web of human knowledge and so an investigation there may provide answers. It is well known that Wikipedia has underlying structures, [3] therefore, by a large-scale graphing of articles and links, we can determine if there exist any trends in our spreading of the facts. This process can be done to arbitrary scale as well as for COVID-19 articles - a novel approach to the topic modeling of the pandemic.

Table 1: Summary of statistics

Variable	Definition	Min	Median	Max	Sd
Log cases by episode date	The best estimate for the date of disease onset	6.78	7.59	8.46	0.48
Log B.1.1.7 cases by episode date		0.69	4.96	7.83	1.88
Log B.1.351 cases by episode date		-6.91	0	2.64	3.59
Log P.1 cases by episode date		-6.91	1.61	3.61	4.41
Date		2021-01-01	2021-03-01	2021-04-30	NA
Retail	Restaurants, cafes, shopping, et cetera	-79	-40	-8	10.8
Groceries	Food stores, drug stores, et cetera	-74	-11	31	11.77
Parks	Local/national parks	-40	-1.5	142	37.08
Transit	Subway, bus, and train stations	-83	-63	-45	5.58
Work	Workplaces	-85	-40.5	-9	14.02
Residential	Homes and apartments	6	17	33	4.71

Table 2: Correlation analysis

	Log Cases	Log B.1.1.7	Log B.1.351	Log P.1	Retail	Grocery	Parks	Transit	Work	Residential
Log Cases	1.00	0.62	0.41	0.43	-0.27	-0.15	0.01	-0.49	-0.32	0.41
Log B.1.1.7	0.62	1.00	0.41	0.50	0.06	0.10	0.23	-0.23	-0.17	0.21
Log B.1.351	0.41	0.41	1.00	0.64	-0.05	-0.06	0.03	-0.10	-0.09	0.08
Log P.1	0.43	0.50	0.64	1.00	0.01	0.01	0.19	-0.09	-0.04	0.06
Retail	-0.27	0.06	-0.05	0.01	1.00	0.81	0.20	0.51	0.53	-0.58
Grocery	-0.15	0.10	-0.06	0.01	0.81	1.00	0.23	0.42	0.50	-0.48
Parks	0.01	0.23	0.03	0.19	0.20	0.23	1.00	0.37	0.34	-0.35
Transit	-0.49	-0.23	-0.10	-0.09	0.51	0.42	0.37	1.00	0.59	-0.66
Work	-0.32	-0.17	-0.09	-0.04	0.53	0.50	0.34	0.59	1.00	-0.94
Residential	0.41	0.21	0.08	0.06	-0.58	-0.48	-0.35	-0.66	-0.94	1.00

2 METHODOLOGY

2.1 Google Community Mobility Reports and Incidence for COVID-19 and variants

Incidences of COVID-19 and variants were obtained for analysis from the COVID-19 Ontario website; similarly fetched was the GCMR. [1] [2] Preliminary analysis included a graphing of all mobility data for all public health zones (See figure 1), and all pre/post-processing was done using R in RStudio.

The data set for incidences includes a statistic for episode date (by all and variants), which is a better indicator for case onset and thus was selected as the dependent variable for analysis. The natural logarithm of this was then taken as standard practice for data preparation. The GCMR data was similarly handled by computing the 7-day rolling averages. In general, Google recommends against comparing mobility data because of variations in location accuracy and the meaning of categorization changes from place to place. To remedy this, public health zones with missing data for a sufficiently long period (or the period of analysis) and discrepancies in classification were excluded. (For example, *Haliburton*, *Kawartha* and *Pine Ridge* are not grouped congruently between the GCMR and Ontario incidence data-sets.) As with any data-set, interpretations may be limited, in this case, by the unreliability in the baseline mobility measurement for Ontario. Table 1 summarizes the statistics used.

2.2 Wikipedia graphing

By writing *Python* scripts using the packages *BeautifulSoup* and *Requests*, web crawlers were created to pull all topic links from an article recursively. By parsing a Wikipedia article and selecting for all links to separate articles, we obtain a graph with one node and n links. Repeating this process by visiting a random one of said links grows the graph as desired (or by visiting an entirely random link in case of nullity or error). The results are offloaded from memory through standard note writing techniques.

The software *Graphia* was then used to graph the results. By minimizing the number of edges, cycles can be removed from the data to create what is known as a *spanning forest*. Naturally, articles with more edges will form as earlier branches and a tree is created starting from the initial condition. It is from unusually large branches that we can deduce over-reporting. Not all pages need to be visited since the article title is deducible from other nodes and node titles are never revisited, hence, scraping is linearly complex.

The structure of human thought is, by definition, the branches of Philosophy - Epistemology, Axiology, Logic, and Metaphysics. In mapping Wikipedia, we would expect these, then, to become the main branches in the graph. The article (https://en.wikipedia.org/wiki/Philosophy) was used as the initial condition since it is known that a recursive clicking of the first link

Table 3: Regression analysis - Coefficient estimation and P-values for COVID-19 and variants (Note: results not scaled to [-1,1])

Variable	CE All	P	CE B.1.1.7	P	CE B.1.351	P	CE P.1	P
Date	-0.0065	0.0912	0.0391	<0.001	-0.0013	0.665	0.0004	0.9142
Date^2	0.0003	0.0025	0.0002	0.1407	-0.0003	<0.001	-0.0008	<0.001
Parks	-0.0148	0.0322	0.0172	0.0952	0.003	0.56	0.0081	0.1909
Parks^2	0.0001	0.0071	0.0001	0.3966	0.0001	<0.001	0.0001	0.0617
Work	-0.41	0.0432	-1.0602	<0.001	-0.1243	0.4171	-0.3282	0.0724
Work^2	-0.0054	0.014	-0.0139	<0.001	-0.002	0.2361	-0.0041	0.0395
Retail	-0.0438	0.5932	-0.2132	0.0819	-0.1233	0.047	-0.3582	<0.001
Retail^2	-0.0014	0.1983	-0.0066	<0.001	-0.0022	0.0085	-0.0058	<0.001
Grocery	0.1122	0.0115	0.0411	0.5347	0.0946	0.0049	0.1601	<0.001
Grocery^2	0.0014	0.4176	0.0088	<0.001	0.0031	0.016	0.0072	<0.001
Transit	-0.196	0.0145	-0.1593	0.1829	0.0031	<0.001	0.3855	<0.001
Transit^2	-0.0017	0.0122	-0.0004	0.6842	0.0015	0.0032	0.0027	<0.001
Residential	-1.4125	0.0022	-3.8185	<0.001	-0.4198	0.2279	-0.6897	0.0966
Residential^2	0.0528	<0.001	0.1226	<0.001	0.0211	0.0382	0.032	0.0084
R^2, R^2 Adjusted , σ	0.658, 0.653, 2.582		0.45, 0.442, 3.856		0.913, 0.912, 1.952		0.872, 0.87, 2.327	

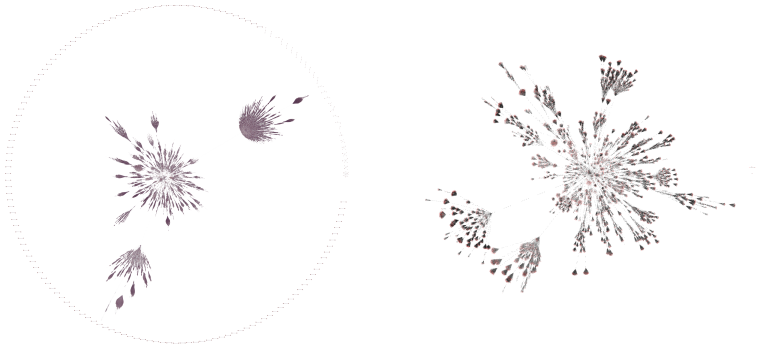

Figure 2: Left: *Graphia* map of *Wikipedia*. In the south-west and north-east 'bursts' the main nodes are *Latin* and *History* respectively. Right: COVID-19 Wikipedia map

of any article leads to this page 97% of the time. [3] This process was repeated for COVID-19 with the initial condition 'https://en.wikipedia.org/wiki/COVID-19'. For the latter, depth was limited to $d = 3$ (where d is the number of recursive steps) with no completely random visits.

3 RESULTS

3.1 Statistical Analysis

Table 2 demonstrates that mobility indicators correlate well with each other. We only consider 'all cases' for the variable selection in multivariate analysis (although it is performed on all variants). The obvious choice is 'Grocery' and 'Residential' because its opposing correlation with all cases and the negative correlation between each other increases the information received from multivariate analysis. It also represents two opposing ideas: mandatory places of high and low risk of COVID-19 contraction.

Table 3 demonstrates the results of the regression analysis. Covariates in the regression are of $O(2)$ and most are significant for at least P<0.05. The positive signs in Grocery^2 and Residential^2 demonstrate non-linear growth, and the +/- signs in the normal variable demonstrate the direction respectively.

We then visualize these results based on predictions from the multivariate model. Figure 3 shows the gradient surfaces of 3 days starting from the announcement of the lockdown and the (2) weeks after. The hyperbolic plots for the given date(s) indicate the expectation of increased case incidence as functions of Grocery and Residential mobility. We note differing depth for the different variants (See Figures 5,6,7). We notice trends in the tracking of various public health units such as Hamilton and Toronto, and interestingly, the segregation of York.

3.2 Biases in the representation of Wikipedia in its entirety and for COVID-19 articles

After scraping ≈ 95.39% of Wikipedia article titles (6003745), we see a clear bias towards the articles *History* and *Latin*. By dividing the number of pages visited by the number of nodes, we find the average number of article links to be 113. The center is the initial condition.

For the latter graph, no major biases are observed. The articles 'Social Distancing' and 'World Health Organization' are observed to be medium-sized (See Figure 4).

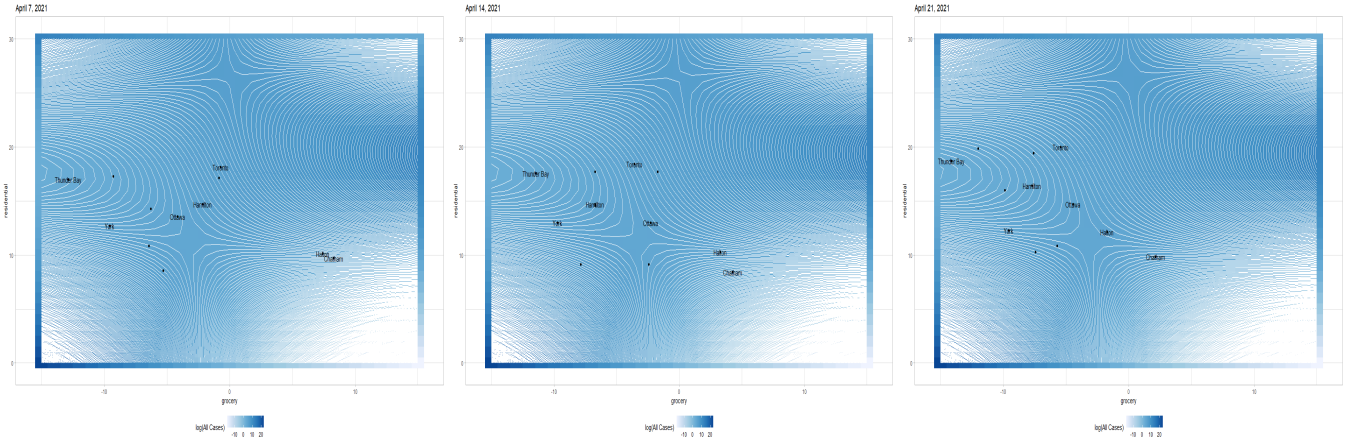

Figure 3: Predicted case incidence gradient vs mobility for the period after lockdown

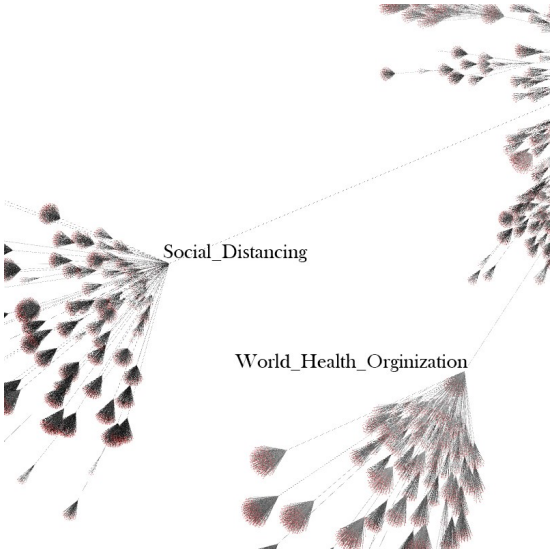

Figure 4: COVID-19 graph close-up

4 DISCUSSION

We observed that Grocery and Residential based mobility offset each other since, in some sense, they are inherently related. Further, we can realize public health zone relatedness by examining such static life dependencies. This is what appears to be the case for Hamilton and Toronto and indicates valid correlations for the mobility predictions. The result observed may be because workers from those regions often live in the other respective region. Despite the caveat of reduced policy-making interpretability from the somewhat inaccurate baseline measurement, even with this, it is clear that intuitively and statistically, the province-wide shutdown was rash.

As for the topic modelling results, unsurprisingly, articles related to the transmission dynamics became rooted in the tree. Curiously the WHO was considered equally important, which suggests that the politics of the pandemic are as relevant as its cause.

REFERENCES

[1] [n.d.]. All Ontario: Case numbers and spread. https://covid-19.ontario.ca/data. Accessed: 2021, April.
[2] [n.d.]. Google Community Mobility Reports. https://www.google.com/covid19/mobility/. Accessed: 2021, April.
[3] [n.d.]. Wikipedia:Getting to Philosophy. https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy. Accessed: 2021, January 30.
[4] Antonio Paez. 2020, June 29. Using Google Community Moblilty Reports to investigate the incidence of COVID-19 in the United States. *Transp. Findings* (2020, June 29).

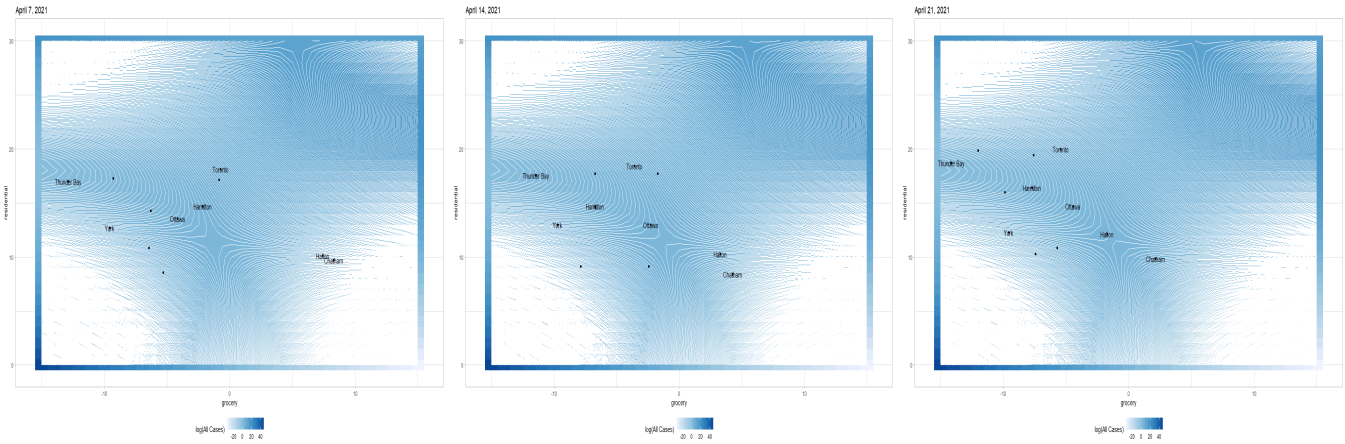

Figure 5: Predicted B.1.1.7 incidence gradient vs mobility for the period after lockdown

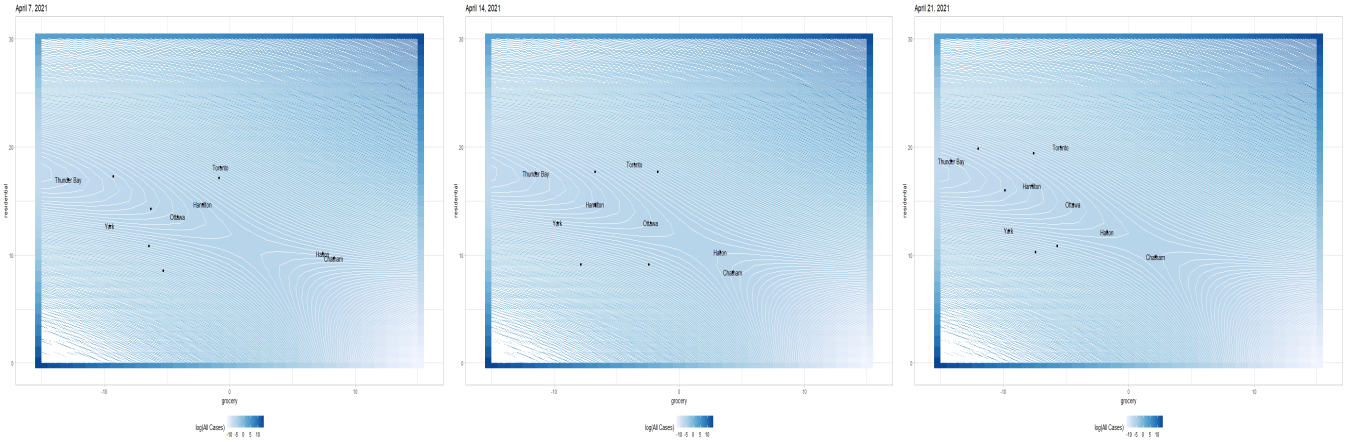

Figure 6: Predicted B.1.351 incidence gradient vs mobility for the period after lockdown

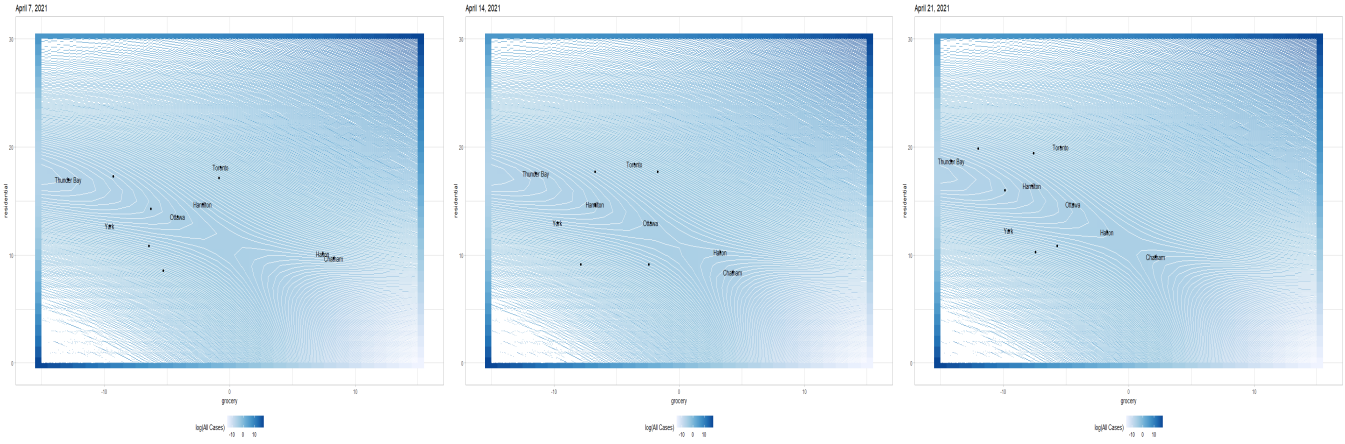

Figure 7: Predicted P.1 incidence gradient vs mobility for the period after lockdown