How information about COVID-19 travels across the public, media science community and policy makers and affects public behaviour

1 Project overview

Information related to health risks and healthy behaviour is typically generated by scientists and distilled into recommendations by public-health agencies. From there it is often transmitted by mass media to the public. Misinformation, generated by careless or irresponsible scientists, pseudo-scientists, or product-marketers can follow a similar route. In the age of the internet, these routes remain important, but the public also has easy direct access to information from public-health agencies, and to many scientific papers, including unvetted preprints. The public also has an expanded ability to interpret, transmit, and amplify messages through social media. More recently still, scientists and agencies have become active on social media as well.

During disease outbreaks of international concern, like pH1N1, WAEO ((fill in better names with dates)) or the current COVID-19 outbreak, this process is both compressed and amplified. The stakes also become higher, because public behaviour directly affects the spread of infectious disease: people who avoid large gatherings may slow disease spread, while people who flee infected areas may accelerate it, for example. Excessive fear of disease spread can have severe economic effects, and may also lead to bias and discrimination against groups seen as linked to the disease, or merely seen as others.

The proposed research will study how information flows between forums, including scientific publications; governmental policies and agency recommendations; and mass and social media – and investigate how it affects public perceptions and behaviours. Our interdisciplinary group will analyze these flows by combining textual and contextual analysis; AI-assisted human-supervised data mining; time-series analysis; and mathematical modeling. We will study how good information competes with misinformation, and look for factors correlated with successful spread of good information. We will gather information on communication and surrogates for behaviour from a wide range of sources. Sources about communication will include agency websites; preprint servers and publicly available scientific journals; major newspaper websites; social-media platforms; and twitter data. Surrogates for public perceptions and behaviour will include twitter data (again); google trends; publicly available records for movie majors box office receipts; publicly available travel data; and information about cancellations and shortages (for example of face masks or pharmaceutical or pseudo-pharmaceutical products) from our textual analysis.

The project will be organized around three Research Questions:

RQ1 How does information (and misinformation) travel between scientists, public-health workers, mass media and social media? (Research sub-area "cultural dimensions of the epidemic")

RQ2 How does communication affect public behaviour and the course of the outbreak? (Research sub-area "public health response")

RQ3 How can scientists and policy-makers evaluate and improve the effectiveness of their communication? (Research sub-area "strategies to combat misinformation")

2 Background

Public-health communication is a balancing act. Officials are often caught between the need to be heard, and the danger of causing panic. This problem is particularly acute in the case of an infectious outbreak, since the presence of a novel pathogen increases both the importance of being heard and the danger that the public will over-react.

In the case of COVID-19, scientists are still scrambling to understand the pathogen's biology; public-health workers are scrambling to decide on the best recommendations and policy decisions given current knowledge at any given time; and the mass media is scrambling to understand the situation and decide how best to communicate with the public.

There are other complicating factors. An outbreak of global concern represents an opportunity for mainstream and peripheral media, and for social-media actors to increase their "clicks" and "likes" and therefore prestige and/or profitability. These motivations work against the balancing act, and instead favor over-simplification and sensationalization.

2.1 Media and Infectious Disease

((mli: This is where I think C's text goes.))

Success of curtailing disease outbreak depends on not only public health system, but also public adherence based on their perception of the disease. Media have influence in swinging public perception of health issues and their estimation of health risk by framing what to present and in what context and in affecting their risk estimation, for example, overestimating rish during SARS outbreak [**?**] or triggering vaccinating panic on a transmission dynamics of influenza [**?**]. It can also be biased against "others" as Western newspapers portrayed Chinese during the SARS crisis [**?**] (https://doi.org/10.1080/01292980500261621). Media is also the tool public health authorities rely on to promote their concerns and recommendations during health risk outbreak. Understanding media effect on disease spread (e.g., media attention increases self-protection) can help enhance epidemic forecasting and preventive measures to slow the disease spread [**?**].

While news media (due to its online virtue) still carry their influence, social media is shaping how we communicate and understand information [**?**]. When social media undoubtedly provided benefits for public health promotion and disease prevention [**?**, **?**], for example, people using twitter for health information were merely likely to be vaccinated [**?**], it also impose potential threats, such as discrimination, disinformation and misinformation [**?**, **?**]. Since the initial reports of cluster of acute severe respiratory disease (COVID-19) and the potential for global spread, there has been widespread discussion and dissemination of information through social media [**?**]. Twitter is a real-world disease outbreak-reporting source, and could report case outbreak ahead of official reports. [**?**]. It can be used to circulate information from credible sources but also as a source of experiences and opinions [**?**]. The nature of some of the information and conversations on social media could be discriminatory and increase panic about the risk of infection. The feeling of anxiety in the population, in part can be assessed by increased demands and shortages for face masks and hand sanitizers even in countries such as Canada where only x cases confirmed to date.

(CS: we need a few setences about policy makers and google trend impact and roles to tie this all together)

((mli: maybe make a plot using google trend for the keyword search "coronavirus" and HuBei time series))

3 Methods and feasibility

3.1 Data and Sources

Media data from November 2019 (before when the virus was officially acknowledged to ensure no related news was overlooked) till June 2020 (when the outbreak is expected to temper down) will be collected. Based on the virus impact (geographically, socially and epidemically) news media with top circulations and online access will be selected from Singapore, Taiwan, Hong Kong, China, Canada and USA and England. In addition, we plan to gather Sing Tao Daily, and World Journal, the most circulated Chinese news papers in Canada and United States respectively. The chinese news will add dimension to cultural and social perception of Chinese immigration population in North America.

We will use Lexis-Nexis search engine to extract news. OriProbe Information Services `https://www.oriprobe.com/peoplesdaily.shtml` provides access to the archive of People's Daily of China. Key words used for searching is attached (Becky's figure)

Twitter data (countries and region covered?) will be collected by (python or AI or purchased? need some details on how to access tweets.)

GoogleTrends accumulate data of search frequency around the world by region and county and by time. It reflects awareness-related burst of searches [**?**, **?**]. We use it as an index to reflect the concern and interest the public expressed during the outbreak.

(CS: we can delete this part or move it to coding book) key words search for Google trend, and official public health websites (discussed below) include coronavirus, 19-nCov, COVID-19, influenza, 武疫(Wu Virus)，武病毒(Wuhan Virus)，武(Wuhan)，武肺炎(Wuhan pneumonia)，肺炎(pneumonia)，冠状病毒(coronavirus)，新型冠状病毒(novel coronavirus)，武新型冠状病毒(Wuhan novel coronavirus).

(CS: someone fill in how we will collect and analyze the trends data)

The epidemic findings from science community is part of the information the public, media and international community rely on. They play a powerful role during public health crises due to the time urgency with which they can disseminate new information, accurate or not. (ref: Early in the Epidemic: Impact of preprints on global discourse of 2019-nCoV transmissibility `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3536663`. We can also check findings of preprint (speedy information delivery, lack of peer review) in terms of accuracy and how misinformation get circulated based on those findings. (see examples(`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3536663`))

Science publications of COVID-19 (peer-reviewed and preprint) will be extracted by searching Google Scholar, arXiv, bioRxiv, medRxiv, and SSRN, , PubMed, Embase, Medline, and OVID). We will construct themes of the studies and results. Our team will evaluate the impact of the publications and compare to whether it is cited, circulated and twitte/re-twitte by the other forums.

In addition to the above data sources, we will summarize reports, guideline and recommendations on the official health websites: WHO's official coronavirus website , the official

centres of disease control and prevention in the studied countries (e.g., US CDC , National Health Commission of PRC (http://en.nhc.gov.cn/index.html), PHAC, Canada. It is noted that each public health instituted mentioned above launched special pages for this outbreak.

We will collect data relating to the outcome of the COVID-19 outbreaks: travel, movies (`https://www.boxofficemojo.com/?ref_=bo_nb_di_mojologo`), restaurants in Taiwan, Singapore, H.K., Canada and USA; or should we focus on North America and on travel and movie box office (`https://www.boxofficemojo.com/?ref_=bo_nb_di_mojologo`)?

Daily case report data: Daily updates from the Provincal Health Commission of HuBei and outside of HuBei are avalible online. These updates contains information on up-to-date cumulative cases, new cases per day, number of confirm case hospitalized, severity categorizations and deaths.

Final stage of our data analysis, we will plot google trends data, the daily infected cases and fatality and frequency and main themes of media coverage and tweets, we expect to construct a proxy indicator for information dissemination and public reaction to the outbreak. Furthermore, we hope to uncover the communication flow among the four forums.

3.2 Analysis of Data

Data Mining

AI-assisted human-supervised data mining (CS: to fill in ideas on how AI will analyze tweets)

Textual analysis of news media and Twitter will mainly based based on a coding book and implemented by AI. The coding book should list themes and frames for the analysis, which will be categorized based on a pilot study by trained coders supervised by the team after the proposal is funded. The basic structure of the themes will outline theme, information types (e.g., news, information and misinformation based on sources verification), who, where, theme tone and when. (see attached [Basic Coding frame](https://github.com/cfshi/coronavirus/blob/master/cod

- key words search for Google trend analysis, and official public health websites: coronavirus, 19-nCov, COVID-19, influenza, (Wu Virus), (Wuhan Virus), (Wuhan), (Wuhan pneumonia), (pneumonia), (coronavirus), (novel coronavirus), (Wuhan novel coronavirus)

Time-series analysis Google trend, reported daily reports, patterns with policy change/case definitions. (CS: A couple of sentence on how we can hack it)

Contextual analysis Contextual analysis will be studied by the researchers. Media content and information are often socially and politically cultivated []. We will implement contextual analyses of the quantitative findings and elaborate how information is framed and disseminated in this uncertain circumstances into context. Textual analysis using AI is reliable and efficient in terms of data mining. However, it can be insufficient in understanding what contents connote. For example, meanings signified in the public reactions to shortage of face masks in Toronto is likely different from the areas where are heavily impacted with COVID-19 epidemically. Without referring to circumstance, we will miss what the results

signify and its impact on the public. To amend this issue, we will conduct contextual analysis [] to elaborate the results of textual analysis of tweets and news and uncover its social and cultural perspectives of the text meanings.

AI-assisted human-supervised data mining Textual analysis of news media and Twitter will mainly implemented by AI based on a coding book. The coding book should list themes and frames for analyzing the texts, which will be categorized based on a pilot study by trained coders supervised by the team after the proposal is funded. The basic structure of the themes will outline what (e.g.), in addition to information types (e.g., news, information and misinformation/fake news based on sources: verified, anonymous, speculation), who (e.g. active vs. passive subject), where (e.g., country, region and city), theme tone (e.g., negative, positive, neutral) and when (date and before vs. after WHO's update, for example). (See attached [Basic Coding frame] for examples of themes.) [[JD: We don't want links in the proposal. Should we try to make the codebook into an attachment?]]

- key words search for Google trend analysis, and official public health websites: coronavirus, 19-nCov, COVID-19, influenza, (Wu Virus), (Wuhan Virus), (Wuhan), (Wuhan pneumonia), (pneumonia), (coronavirus), (novel coronavirus), (Wuhan novel coronavirus)

(Need sentences on how AI will analyze tweets) We will apply AI techniques to Twitter stream data based on the coding book to analyze tweets and parse out themes and to apply algorithm to identify false information through topic modelling. We will also track evolution of information content and varsity over time and across geographic areas.

Time-series analysis Google trend, reported daily reports, patterns with policy change/case definitions.

Mathematical modeling Modelling: (Can we do this?: incorporating media coverage, public and institutional reaction and report (which can be based on modelling results by scientific community and WHO etc), time travelling and spread network (geographical)

3.3 deliverables

- software

- Communication: Writing papers is certainly one, but we should be tweeting and etc. What are communication platforms we should be communicating in?

4 Research Setting & Personnel

The principal applicants have a long history of influential research on disease outbreak:

The research will principally take place at McMaster University. Nominated principal applicant Dr. David J.D. **Earn** (6 hours per week) led the creation of the International Infectious Disease Data Archive [33] and has expertise in gathering and curating infectious disease data, and in dynamical modelling. Principal applicant Dr. Jonathan **Dushoff** (5 hours per

week) is an internationally recognized expert in infectious disease modelling, has extensive experience with statistical frameworks for fitting models to data, and has been involved in the Ebola challenge and other forecasting projects. Co-Applicant Dr. Benjamin **Bolker** (5 hours per week) is a highly accomplished ecological statistician with extensive experience in spatial-dynamical modeling, statistical modeling and statistical software. Co-Applicant Dr. Chyun-Fung **Shi** (40 hours per week), post-doctoral researcher, has ... Co-Applicant Dr. Michael **Li** (5 hours per week), post-doctoral researcher, has focused his research on epidemic forecasting and is experienced working with large databases.

((mli: fill in the rest later))

4.1 Collaborators and Knowledge Users

((mli: fill in later))

5 Research Time line

5.1 Year 1

5.2 Year 2

5.3 Potential Outcomes

Transfer findings to peer-reviewed publications. In addition to a grand one paper, we plan to transfer findings into subgroups, such as by affected country (North America, Asia-Taiwan, H.K. and Singapore), types of media (news media and social media (twitter)).

- Potential to contribute to the global response to COVID-19
- Social and policy countermeasures and Global Coordination Mechanism

6 Challenges and Mitigation Strategies

Social media in China is to included due to data availability. Social media such as twitter and Facebook are barred in China and WeChat does not share their database. Yet, there are information and coverage about the outbreak inside China in the media. We will categorize people's reaction and understanding of COVID-19 on the news content or the tweets to supplement our results.

Data curation We will go back and document clearly all the policy changes and case defintions.

- data collected from National Health Commission
- Figure out how to use the data effectively (e.g. we are not using death, severity categorizations, number of tested, number of positive, and etc)
- Case definitions
- Media content

- Social media platforms
- Google trends

Analysis/ Pipelineing/ mainstreaming Info delays, misinformation and miscommunications

Communication how people are interpreting

7 Conclusion/summary

7.1 Note

We want to emphasize our bilingual (English and Chinese) background, which is essential in this study.