Routes of communication: understanding the information flows that shape public and official reaction to COVID-19

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 2003 SARS, 2009 H1N1, 2014 Ebola, 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 identified 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 dynamical 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; and social-media platforms. Surrogates for public perceptions and behaviour will include data from open social-media platforms twitter and Weibo; google trends; publicly available box-office information for movies and major sporting leagues; and publicly available travel information; 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 (RQs), all based on the Social and policy countermeasures research area from the funding opportunity:

RQ1 How does information (and misinformation) travel between scientists, public-health workers, mass media and social media? (*based on funding opportunity research sub-area*: "cultural dimensions of the epidemic")

RQ2 How does communication affect public behaviour and the course of the outbreak? (*research sub-area*: "feasibility and effectiveness of 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 disease 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.

It is known that traditional media can strongly influence public perceptions, creating fear by overestimating risk during the SARS outbreak [?] and the influenza pandemic [?], or feeding into bias – e.g., anti-Chinese bias during both the SARS crisis [?] and the current outbreak. Media is also the tool public health authorities rely on to promote their concerns and recommendations during outbreaks. Understanding media effects on disease spread (e.g., media attention increases self-protection) can help enhance epidemic forecasting and preventive measures to slow the disease spread [?].

While traditional news media (including online presence) remains influential, social media plays an increasingly important role in shaping how we communicate and understand information [?]. Social media can play a positive role spreading good information, [?, ?, ?], but may also spread misinformation and feed bias [?, ?]. Since the initial reports, a 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 [?, ?, ?].

3 Methods and feasibility

3.1 Data

Science We will develop systematic search and screening strategies to extract relevant peer-reviewed publications from Google Scholar and PubMed. To account for the strong influence of preprints early in the epidemic [?], we will also include preliminary scientific findings posted on medRxiv through 31 March 2020. We will index these papers and track their appearances in mass media and social media; we will also track which of the preprints are published after peer review.

Public health recommendations We will collect and analyze reports, guidelines and recommendations available from the World Health Organization and from the central disease control agency of each of our focal countries. It is worth noting that all of these agencies have launched special COVID-19 web pages.

Mass media We will use the NexisUni search engine (via McMaster University) and OriProbe Information Services to collect articles relevant to the outbreak, going back to the outbreak start in December 2019, and continuing throughout the grant period. We will focus on the top English- and Mandarin-language newspapers (taking both circulation and online access into account) from Canada, China, England, Singapore, Taiwan, and the USA. We will include the top Mandarin-language newspapers in both Canada and the USA.

Social media We will efficiently collect data from Twitter and Weibo by purchasing API access, using data going back to November 2019 – before the epidemic started – to give a baseline for comparison.

Public response Twitter (and Weibo) data will give us information not only on information flows, but also on public interest, attitudes and topics being discussed on the social media. We will also probe public interest and concern using publicly available data from GoogleTrends, which tabulates frequency of searches (by search times and topics) in various regions across the world [**?**, **?**]. Economic data relating to the outcome of the COVID-19 outbreaks will be gathered as reference for contextual analysis (see below): for example, travel, movies box offices, cancellation of public events.

3.2 Analysis

Textual analysis To investigate how information/misinformation travels and how communication affects public responses (e.g., attitudes), we will use state-of-the-art machine learning and natural language processing (NLP) techniques. We have two directions. First, to investigate how information travels, we plan to develop codebooks to manually annotate random samples of articles/messages from scientific papers, government recommendations, mass media and social media. Codebooks will contain both themes and frames relevant to our analysis. For example, correct or incorrect information (misinformation or not?) will be annotated. Then, using the annotated data as training data, supervised machine learning algorithms such as logistic regression, Support Vector Machine, and neural network will learn patterns of textual and content features to predict the label(code?). Second, to study public responses, we will leverage NLP techniques such as aspect-based sentiment analysis and topic modeling. For example, ABSApp, a state-of-the-art aspect-based sentiment extraction tool, allows extracting important aspects of the target(?) situation and detecting sentiment towards the aspects, with relatively less human involvement. Latent Dirichlet Allocation (LDA) and its variants, popularly used topic modeling techniques, help identify topics in large texts. These approaches will allow us to perform a large scale study (using large data). For all the textual analysis processes, computer scientists will iterate collaboratively with coders and subject-matter experts to build codebooks that are consistent with study aims, and to interpret findings from applying AI algorithms. The findings would be shared with public-health practitioners to assist with counter-messaging strategies.

Time-series analysis We will use cross-correlation analyses to look for indicators that information is moving from one communication forum to another; that events (like disease spread or public behavior) are affecting communication; or that communication is affecting events. Cross-correlation analysis is complicated and prone to false-positive results. Importantly, therefore, we will be able to use the cross-correlation analysis to generate hypotheses that can be checked by more detailed textual analysis. For example, if we hypothesize that tweets about fatalities are being driven at a certain time and place by mass media, or by government policies, we can sample from those tweets and examine them for detailed information or citations; if we hypothesize that a trend in self-isolation is driven by social media, we can search for mass media stories that interview people about their motivation. The ability to compare large-scale trends with detailed texts should amplify our pattern to detect and confirm patterns.

Dynamical modeling Dynamical modeling provides the link between individual events and emergent phenomena. We will make a range of simple dynamical models to further probe our time-series results by asking what mechanisms may underlie our observed connections, and what these connections might imply for the future. Dynamical models will allow us to explore hypotheses about what factors affect behaviour, and also to explore new hypotheses about how changes in behaviour are likely to loop back to disease transmission or to panic responses that might lead to shortages or to impacts on regional economies or the global economy. **((naveed: This is not clear to me what we are trying to model here and how this will link back with the other themes.))**

Synthesis Effective health communication, including dissemination of good information and countering misinformation, is key to outbreak management [**?**] and consistent recommendation [**?**, **?**]. We will use techniques from content analysis to combine results from our analysis techniques above to formulate hypotheses about what factors lead to effective communication. In particular, we will identify cases where good information did or did not out-compete bad information. When information from public-health agencies spreads effectively we will also evaluate our behavioural proxies to ask when it led to a calibrated reaction from the public (as opposed to over- or under-reaction). We will make use of previous studies to evaluate [**?**, **?**, **?**, **?**] and develop [**?**, **?**, **?**] strategies for effective health communication.

3.3 Applications

Real-time identification and response We will combine our analysis results to identify examples of misinformation spreading well; good information spreading poorly; and public over- and under-reaction. We will work directly with team members at BCCDC, knowledge users at PHAC and their associates to develop and evaluate messaging strategies. We will also share information about spreading misinformation with public-health workers and the public, through twitter, blog posts and a dedicated web site.

Academic outreach We will share our methods and results through peer-reviewed papers and academic conferences.

Software and data To the extent possible, all of the software developed for this project will be based on open platforms (primarily python and R). All software will be shared publicly via version-control repositories. In particular, the tools we develop to assist textual analysis have the potential to find a wide audience, but we will also share software for applied time-series analysis and dynamical modeling.

4 Research Setting & Personnel

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 and has expertise in gathering and curating infectious disease data, and in dynamical modelling, including modelling the influence of individual decision-making on epidemic dynamics. 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 are integrated around accomplished multidisciplinary researchers with extensive experience in their respective fields. Dr. Chyun **Shi** (40 hours per week) is accomplished social-scientist with extensive experience in social and health behaviour and applied experience in both journalism and advertising. Dr. Jung Hui **Yeh** **((mli: fill in for me))**. Dr. Giuseppe **Carenini** (5 hours per week) and Dr. Hyeju **Jang** (40 hours per week), (to be hired as a post-doctorial researcher), are accomplished computer scientist with extensive experience in artifical intelligence (AI) and computational linguistics. Dr. Mark **Loeb** M.D., M.Sc. (3 hours per week) is an infectious disease clinician and epidemiologist with broad experience in practical public health issues. 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. Dr. Michael **Li** (5 hours per week) has focused his research on epidemic forecasting and is experienced working with large databases.

In addition, we have an extensive collababorate team supporting and enchance the research, providing expert opinions on results, knowledge translation and communication.

Dr. Naveed Z. **Janjua** (5 hrs week) at the BCCDC, leads data and analytic services and was involved in the 2009 H1N1 pandemic response; during pandemic, lead or contributed to studies on immuno-epidemiology of pandemic H1N1, household transmission, modelling, effect of prior seasonal vaccine receipt on pandemic H1N1 infection risk and pandemic vaccine effectiveness.

Dr. Xing Peng Jiang **((mli: Help me fill this in))**.

At Public Health Agency of Canada, Chief Science Officer, Dr. Pascal **Michel** role is to understand high-level inter-relationships between various programs and priorities. Epidemics, emergencies and disasters often bring pressures on various organizations to produce timely information to guide decisions making.

Dr. Nai Rui **Chng** is a Political Scientist with an interest in the development and evaluation of complex interventions in health, social and environmental policy domains. He is a versatile qualitative researcher who works in high, middle and low-income countries.

Our team is ready to respond rapidly, and in fact has already started doing so. Co-applicant Jang has begun to collect twitter data, and co-applicant Li has been working on curating data about the epidemic. Co-applicant Shi has interim funding to begin working on gathering google trends and media data as soon as this proposal is submitted.

We are also already in touch with public-health response officials through BCCDC and

KFL&A, and will reach out through our collaborators at PHAC as well to provide information directly of use.

5 Challenges and Mitigation Strategies

This proposal is ambitious and will meet with unexpected obstacles. Our main strategy is to actively foster open communication between members of this very diverse team. We are working on clear task definitions, so that everyone is able to move forward and also clear lines of communication so that we are able to advise and make use of each others' work.

China is at the center of the outbreak, but information from China is not always open. Chinese use of twitter and google differs sharply from most of the rest of the world. Team member Jiang is based in Wuhan, and will help us navigate some of these difficulties. We will interpret Chinese data with care, and not put analysis of China at the center of our project. Importantly, we also have team members familiar with Singapore and Taiwan, more open societies that have also felt strong social effects from the outbreak.

The future of the coronavirus outbreak is unpredictable. **((jd: We will be flexible. How?))**

References