Tutorial Proposal for SBP-BRIMS 2022

# Multi-Level Analysis of Online Harm Networks

Joshua Uyheng* and Dr. Kathleen M. Carley
CASOS Center, Institute for Software Research, Carnegie Mellon University
<juyheng@cs.cmu.edu>

## Tutorial Objectives

By the end of this tutorial, participants should be able to do the following:

- define online harms such as hate speech and trolling through a network lens;
- explain individual-level, group-level, and conversation-level measures used to analyze online harm networks; and
- run the pipeline of tools required to execute the ORA Online Harms Report and interpret its findings.

## Introduction

Recent years have exposed the vulnerability of online social networks to various forms of harm. Social media platforms' capacities for large-scale communication have been weaponized by organic and coordinated actors alike to spread hate and disinformation. Such online harms have had profound, multi-level consequences, including the well-being of individuals, relationships within and between groups, and on society as a whole. However, prevailing approaches to assess online harms have largely focused on the sheer volume of harmful messages and actors. This divorces them from their interlinked, network context.

In this tutorial, we discuss a general, multi-level framework for understanding online harm networks. Key forms of online harm will be discussed, and their individual-level, group-level, and conversation-level impacts will be explained to highlight the importance of multi-level analysis. A hands-on case study activity will then be conducted with anonymized data to demonstrate an application of this framework through ORA-Pro.

## Data and Methods

An anonymized dataset of tweets from the 2020 United States presidential election will be used to demonstrate the full pipeline of tools required to execute the ORA Online Harms Report. The dataset contains one million tweets and the standard Twitter metadata required to construct agent-by-agent, agent-by-hashtag, tweet-by-tweet, and tweet-by-hashtag networks.

Running the ORA Online Harms Report will require enrichment of the dataset with information regarding the harmfulness of tweets and agents. We will use Netmapper to obtain direct measures of harmfulness in the form of abusive language, then the CASOS Hate Speech Detector and TrollHunter to obtain further predictions of hate speech and trolling. We will additionally use BotHunter to obtain account-level bot predictions. Using all these inputs, we will finally run the ORA Online Harms Report and interpret its findings.

## Organization of Session

1. (15 minutes) Conceptual explanation of online harm networks and a multi-level framework of analysis
2. (15 minutes) Data enrichment for case study dataset with Netmapper and BotHunter
3. (20 minutes) Running and interpreting the ORA Online Harms Report for the case study
4. (10 minutes) Q&A