Non-parametric Information Flow Estimation in Social-Media News

Tobin South

Feb 2021

Abstract

News media has long been an ecosystem of creation, copying and critique, where news outlets break stories and add new information and opinions to ongoing stories. Understanding the dynamics of how news information is created and spread is important to accurately ascribe credit to influential work and understand how societal narratives develop. These dynamics can be modelled through a combination of information-theoretic natural language processing and networks using large quantities of textual data.

In this thesis, new comparative techniques are developed to estimate textual information flow between pairs of news-media outlets. To achieve this, Chapter 3 outlines the collection and cleaning of data sourced from the Twitter accounts of news-media organisations collected for the duration of 2019. Chapter 4 introduces two non-parametric entropy estimators and validates their convergence. These estimators are then extended to produce several measures of information flow in Chapter 5, which are compared via simulation models using both synthetic and real text data. The resulting best estimator produces a reliable measure of textual information flow that captures all aspects of grammar and word choice in its calculation.

Resulting information flows are constructed into a network of news-media organisations, which Chapter 6 uses to examine approaches of ranking influence using examples from centrality, sport ranking and network topology. The interconnected nature of the information flow ecosystem proves resistant to simplification, demonstrating the implicit complexity in the flow dynamics.

In total, this work provides a new methodology for examining the information transmitted between content producers in any connected system of natural language, a toolkit with applications to the many networked discourses of our online world.