Physica A 465 (2017) 289–296

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Modeling the infectiousness of Twitter hashtags

Jonathan Skaza [a,*], Brian Blais [b,c]

[a] Department of Biostatistics, University of Michigan, United States
[b] Department of Science and Technology, Bryant University, United States
[c] Institute for Brain and Neural Systems, Brown University, United States

HIGHLIGHTS

- Trending hashtags fall into two groups: slightly infectious and very infectious.
- An automated approach to quantify the trendiness of hashtags is developed.
- The basic SIR model appears to be able to adequately capture the dynamics of a trending hashtag.

ARTICLE INFO

Article history:
Received 25 February 2016
Received in revised form 29 June 2016
Available online 21 August 2016

Keywords:
Twitter dynamics
Trending
SIR
SIRI
MCMC
Information diffusion

ABSTRACT

This study applies dynamical and statistical modeling techniques to quantify the proliferation and popularity of trending hashtags on Twitter. Using time-series data reflecting actual tweets in New York City and San Francisco, we present estimates for the dynamics (i.e., rates of infection and recovery) of several hundred trending hashtags using an epidemic modeling framework coupled with Bayesian Markov Chain Monte Carlo (MCMC) methods. This methodological strategy is an extension of techniques traditionally used to model the spread of infectious disease. Using SIR-type models, we demonstrate that most hashtags are marginally infectious, while very few emerge as "trending". In doing so we illustrate that hashtags can be grouped by infectiousness, possibly providing a method for quantifying the trendiness of a topic.

1. Introduction

Twitter (http://twitter.com) is a popular social networking website that allows users to both send and read messages known as tweets. The social networking site has approximately 320 million monthly active users that produce an average of about 500 million tweets per day [1]. Twitter serves as a place for users to share anything and everything on their minds—news stories, ideas, quotes, lyrics, etc. Users are able to embed hashtags within their tweets by using the hash character (i.e., #). Hashtags are metadata tags which allow tweets containing the text following the hash character to be grouped together. From there, it is possible for users to query certain hashtags to see what is being discussed throughout the site. The Twitter site even contains a panel of trending topics—hashtags and topics that have become very popular in a short period of time.

Hashtags can also prove useful for researchers in need of categorizing or grouping tweets. While it is also possible to filter tweets by words or phrases, such an approach can be problematic. For instance, a researcher interested in exploring the degree of happiness on Twitter may search for tweets containing the word "happy". While this strategy will return

* Corresponding author.
E-mail address: jskaza@umich.edu (J. Skaza).

http://dx.doi.org/10.1016/j.physa.2016.08.038
0378-4371/

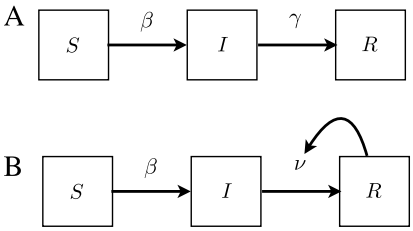

Fig. 1. Infection compartmental models. Shown are the simple SIR model (A) and the more complex SIRI model (B).

tweets from people expressing sentiments such as "I am *happy*", it will also return messages in the category of "I am not *happy*". Sentiment analysis techniques are needed to rectify this issue [2]. Using hashtags to filter tweets hedges against the need to address such concerns; a person who inserts #happy into his/her tweet is likely happy. However, the volume of tweets meeting the specific search criteria will be reduced because not every "happy" tweet, for example, will include #happy. Nevertheless, to avoid problems with contradicting sentiments, the present study uses hashtags as a proxy to study the prevalence and popularity of topics on Twitter.

This study attempts to quantify the spread of certain trending hashtags on Twitter with a *systematic* approach. Using the methods described below, one can estimate the rates of infection and recovery for a particular trending topic. Furthermore, with slight data processing, the same methodology can be used in a predictive context. This study provides a brief overview of the existing literature concerning epidemic modeling and its use describing information dynamics on Twitter. Subsequently, we describe the methods that we used to quantify the propagation of trending topics on Twitter. In the process we find two main categories of hashtag dynamics—marginally infectious and very infectious, without much in between.

1.1. Previous work

Mathematical models have been used in the prediction, control, and analysis of epidemic phenomena—most notably, the spread of infectious disease throughout a population—since the advent of the susceptible, infected, and recovered (SIR) model [3]. These types of epidemic models are featured in studies concerning measles [4–7] and influenza [6,8–10], among others. The basic SIR model describes the dynamical process of disease by categorizing members of the population of interest as either susceptible (S), infected (I), or recovered (R), while incorporating rates of infectiousness (β) and recovery (γ). The parameter, β, controls how often the interaction between susceptible, S, and an infected individual, I leads to an infection and the parameter, γ, represents the rate that an infected individual recovers from the infection, moving from I to R. Fig. 1(A) illustrates a model diagram.

Members of the population transition to and from different compartments based on the system of differential equations presented in Eq. (1).

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI/N\\ \frac{dI}{dt} &= +\beta SI/N - \gamma I\\ \frac{dR}{dt} &= +\gamma I.\end{aligned} \tag{1}$$

Although possibly useful for describing the dynamics of Twitter, it may be that the *infection* of ideas does not follow the same structure. Perhaps, as in Ref. [11], the "recovery" from infection is not a passive time-decay but depends on how many have recovered already. In that case, referred to here as the SIRI model, we have a slightly modified set of equations which can lead to a more rapid decrease in the recovery phase as evidenced in Eq. (2).

$$\begin{aligned}\frac{dS}{dt} &= -\beta IS/N\\ \frac{dI}{dt} &= +\beta IS/N - \nu IR/N\\ \frac{dR}{dt} &= +\nu IR.\end{aligned} \tag{2}$$

Here the SIRI parameter ν takes the place of the parameter γ in the SIR model, but with a different interpretation. The parameter γ is simply a recovery rate, whereas the parameter ν controls how often an interaction between those infected, I, and those already recovered, R, leads to recovery. Thus, in the SIRI model, as more individuals recover the faster the rate of recovery. In the SIR model the rate of recovery is fixed.

A relatively new strategy in the field of epidemic modeling is to develop a statistical model for the parameters of the dynamical model. Specifically, we use Markov Chain Monte Carlo (MCMC) simulation to estimate the posterior probabilities of the epidemic model's parameters (e.g., β, γ, and ν) [10,12].