Linking Epidemic Models and Hawkes Point Processes for Modeling Information Diffusion

Quyu Kong
Australian National University & Data61 CSIRO
Canberra, Australia
quyu.kong@anu.edu.au

ABSTRACT

Epidemic models and Hawkes point process models are two common model classes for information diffusion. Recent work has revealed the equivalence between the two for information diffusion modeling. This allows tools created for one class of models to be applied to another. However, epidemic models and Hawkes point processes can be connected in more ways. This thesis aims to develop a rich set of mathematical equivalences and extensions, and use them to ask and answer questions in social media and beyond. Specifically, we show our plan of generalizing the equivalence of the two model classes by extending it to Hawkes point process models with arbitrary memory kernels. We then outline a rich set of quantities describing diffusion, including diffusion size and extinction probability, introduced in the fields where the models are originally designed. Lastly, we discuss some novel applications of these quantities in a range of problems such as popularity prediction and popularity intervention.

KEYWORDS

Information diffusion, Hawkes Processes, Epidemic Models

ACM Reference Format:
Quyu Kong. 2019. Linking Epidemic Models and Hawkes Point Processes for Modeling Information Diffusion. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3289600.3291601

1 INTRODUCTION

Models of information diffusion help us understand large-scale social phenomena online and offline, such as fake news propagation and the transmission of opinions. Among these models, two commonly used classes of information diffusion modeling methods are the epidemic models [3], typically applied in epidemiology for understanding disease spreading, and Hawkes point processes [4], first used for modeling seismic activities. While the breadth of methodology keeps expanding, Rizoiu et al. [5] have shown the equivalence of the two in stochastic intensities when one type of events (so-called recoveries) are unobserved. This opens the possibility of applying tools created in one model class to another. For instance, a technique for analyzing the size of diseases infected population can be derived for presenting how popularity (i.e. the size) of a diffusion cascade is distributed.

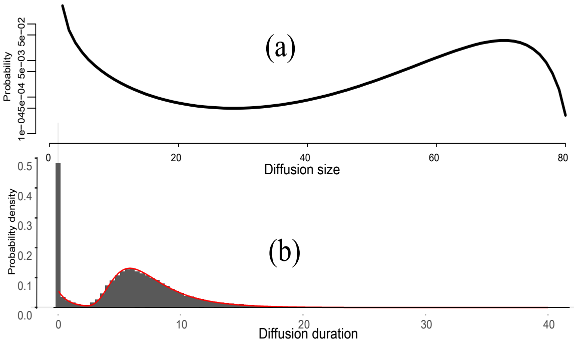

Figure 1: Two inherently unpredictable measurements in social media that have bimodal distributions: (a) Example diffusion size distribution (b) Example diffusion duration distribution.

Meanwhile, epidemic models lead to some important yet under-explored quantities to characterize diffusions. The infected population size, the disease infection duration and the disease extinction probability are several known quantities in epidemiology, whereas only the expected diffusion cascade size is typically analyzed in information diffusion modeling.

In this work, we aim to answer three research questions:

The current connection between epidemic models and Hawkes point processes [5] is valid when individual event influence decays exponentially. A research question then remains open: **how to link the epidemic models and Hawkes point processes with a generalized connection?**

Allen [1] shows how epidemic models quantify the spread of disease in epidemiology in terms of the infected population size, the disease infection duration and the disease extinction probability. However, this level of diversity is absent in quantifying information diffusion which leaves a question open: **how to link quantities derived in epidemic models for characterizing information diffusion?**

How to use the quantities we derived to explain observed phenomena with potential application? Using epidemic techniques, one can compute the distribution of cascade size, which can provide a novel explanation of the unpredictability of popularity. Fig. 1(a) shows the diffusion of a cascade can stop at either a small size or a very large size with high probabilities [5]. Our ongoing work also has shown the existence of similar unpredictability in the time duration of a cascade in Fig. 1(b).

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02.
https://doi.org/10.1145/3289600.3291601

2 BACKGROUND AND RELATED WORK

Epidemic models typically divide population into several compartments and define the transition dynamics among different compartments [1]. The deterministic SIR model is one of the well-studied epidemic models where three compartments (i.e. susceptible, infected and recovered) are introduced and transition dynamics are defined by three differential equations regarding the size of each compartments. A stochastic formulation of SIR models described in [6] is driven by two point processes with correlated stochastic intensity, the infection process and the recovery process.

A *Hawkes process* is a special type of point process which incorporates the idea of self-excitation, i.e. the occurrence of current events will increase the likelihood of future event appearance [4]. The *HawkesN model* proposed by our recent work [5] is a variation of the Hawkes process. It assumes a finite population size as online information tends to diffuse within local communities.

A link between the HawkesN model and the stochastic SIR model is revealed by Rizoiu et al. [5] showing that the HawkesN intensity function is in fact equivalent to the expected infection intensity function of the stochastic SIR model without observing recovery process.

Model quantities: a Markovian technique is shown in [1] for deriving the distribution of the diffusion size and a recursive equation for computing the expected last recovery time of SIR processes in epidemiology. In the Hawkes point processes, on the other hand, the mean diffusion size can be predicted via an additional prediction layer. The branching factor is another known quantity that can be derived via a given equation [4].

3 THE PROPOSED RESEARCH

In this section, we will show some ongoing work and future directions for addressing the open questions.

A Generalized Connection The epidemic-Hawkes connection only has been shown for the exponential individual decay, i.e. the influence from an infected individual decays exponentially. We aim to generalize the connection for arbitrary decay functions.

We note that the decay function of the HawkesN model is in fact the CCDF (Complementary Cumulative Distribution Function) of the recovery distribution an infected individual follows. This leads to a new recovery intensity given an arbitrary decay function for the HawkesN model. This new recovery intensity function defines a new formulation of stochastic SIR models which allows any recovery distribution and, more importantly, the equivalence between the HawkesN model and the SIR model is valid for any decay function.

Diffusion Quantities Exploration Table. 1 lists the models we discussed in Sec. 2 and some known quantities: *diffusion size* is the number of total events in a diffusion cascade; *diffusion duration* is the time duration of a diffusion cascade which is equivalently the last infection event time in a SIR process; *last recovery event time* is the last event time of a stochastic SIR process; *branching factor* is the expected number of events triggered by a new occurrence; *extinction probability* indicates the probability of a Hawkes process will eventually stop.

We first aim to expand our quantity table horizontally by exploring other quantities for information diffusion modeling. Our work is to formalize these quantity definitions and theoretical inferences. one example quantity is *diffusion speed*: a temporal indicator about how fast a cascade size will grow in time is missing. As the popularity and diffusion duration are known, it is intuitive to explore this concept.

Table 1: Quantities and Equivalence across Models

√ : the quantity exists and inference method known.
? : the quantity exists, but the inference method is unknown.
/ : the quantity does not exist

	SIR	Hawkes	HawkesN	...
Diffusion size	√	?	?	...
Diffusion duration (i.e. Last infection time)	√	?	?	...
Last recovery time	√	/	/	...
Branching factor	/	√	√	...
Extinction probability	/	√	/	...
...	...	...	...	...

On the other hand, we plan to complete the missing inference methods for diffusion quantities in Table. 1. Our recent work shows that one is able to extend the expectation of extinction time to the whole distribution and also derive the distribution of the diffusion duration (i.e. the last infection event time in SIR processes), inspired by the notes in [2].

Diffusion Quantities Application While exploring diffusion quantities, we present our work on the exploitation of the existing diffusion quantities in applying them to solving real-world problems.

Prediction Predicting popularity is one of the typical applications information diffusion models where cascades are modeled and the expected numbers of future events or total events are predicted [4]. With the new diffusion properties, however, we can conduct a more thorough prediction. Popularity prediction and diffusion duration can be achieved in both the expected value and the distribution. This allows one to describe the future prediction on a given cascade in terms of its size and duration with a confidence level.

Intervention The models and properties enable one to intervent cascade diffusion and control its propagation in several ways: (a) from the distribution of popularity and diffusion duration, we can analyze the best time and the amount of promotions should be invested to a diffusion to maximize the outcome, which helps content producers make strategic promotion decisions beforehand; (b) we can also integrate the models with control theory or reinforcement learning models to produce real-time promotion strategies.

REFERENCES

[1] Linda J. S. Allen. 2008. An Introduction to Stochastic Epidemic Models. In *Mathematical Epidemiology*. Springer, Berlin, Heidelberg, Chapter 3, 81–130.

[2] Miranda Holmes-Cerfon. 2017. Lecture notes in Applied Stochastic Analysis.

[3] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. 2016. Exploring limits to prediction in complex social systems. In *WWW'16*.

[4] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. 2016. Feature Driven and Point Process Approaches for Popularity Prediction. In *CIKM '16*.

[5] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. 2017. SIR-Hawkes: on the Relationship Between Epidemic Models and Hawkes Point Processes. In *WWW'18*.

[6] Ping Yan. 2008. Distribution Theory, Stochastic Processes and Infectious Disease Modelling. In *Mathematical Epidemiology*, Wu J. Brauer F., van den Driessche P. (Ed.). Springer, Berlin, Heidelberg, Chapter 10, 229–293.