AFRL-RI-RS-TR-2021-114

STUDYING THE SPREAD OF FAKE NEWS WITH POPULATION GENETIC MODELS VIA LONGITUDINAL DATA

UNIVERSITY OF HAWAI'I AT MĀNOA

JULY 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-114 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
ALEKSEY V. PANASYUK
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief, Intelligence
Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2021	FINAL TECHNICAL REPORT	MAY 2019 – APR 2021

4. TITLE AND SUBTITLE

STUDYING THE SPREAD OF FAKE NEWS WITH POPULATION GENETIC MODELS VIA LONGITUDINAL DATA

5a. CONTRACT NUMBER
FA8750-19-2-0027

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

June Zhang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
R2RU

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Hawai'i at Mānoa
2500 Campus Rd
Honolulu HI 96822

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2021-114

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project was foundered under DARPA's AIDA program. It studied the problem of combining dynamic and semantic information in news articles to characterize how information changes over time for fake news detection. The projected collected three unique datasets of news articles from various news outlets on the internet: (1) Jussie Smollett incident (1/28/2019 – 3/14/2019, 1009 outlets), (2) Ukraine International Airlines Flight 752 (1/2/2020 – 2/1/2020, 2136 outlets), (3) COVID news in early 2020 (1/8/2020 – 4/9/2020, 10449 outlets). Semantic triples (subject; predicate; object) are extracted from collected article text each day. These triples correspond to nodes (subject & objects) and edges (predicate) from an RDF (resource description framework) graph. A sequence of RDF graphs represents a corpus of articles covering the same event over multiple days. We considered four features to describe how the structure of RDF graphs change over time: copy, mutate, append, and extend. Analyzing the sequence of RDF graphs, containing both dynamic and semantic information, showed a distinction between articles published on the Jussie Smollett incident by major news outlets vs. celebrity news outlets. We showed that the discrete-time, multivariate Hawkes process can be used to model the dynamics of copy, mutate, append, and extend events.

15. SUBJECT TERMS
Semantic triples, RDF, Dynamic and Semantic Information, Multivariate Hawkes Process

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**ALEKSEY V. PANASYUK**
U	U	U	UU	33	19b. TELEPHONE NUMBER (*Include area code*) **NA**

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

The 'Studying the Spread of Fake News with Population Genetic Models via Longitudinal Data' project was pursued under the AIDA program. It studied the problem of combining dynamic and semantic information in news articles to characterize how information changes over time for fake news detection. The projected collected three unique datasets of news articles from various news outlets on the internet: (1) Jussie Smollett incident (1/28/2019 – 3/14/2019, 1009 outlets), (2) Ukraine International Airlines Flight 752 (1/2/2020 – 2/1/2020, 2136 outlets), (3) COVID news in early 2020 (1/8/2020 – 4/9/2020, 10449 outlets).

Semantic triples (subject; predicate; object) are extracted from collected article text each day. These triples correspond to nodes (subject & objects) and edges (predicate) from an RDF (resource description framework) graph. A sequence of RDF graphs represents a corpus of articles covering the same event over multiple days. We considered four features to describe how the structure of RDF graphs change over time: copy, mutate, append, and extend.

Analyzing the sequence of RDF graphs, containing both dynamic and semantic information, showed a distinction between articles published on the Jussie Smollett incident by major news outlets vs. celebrity news outlets. We showed that the discrete-time, multivariate Hawkes process can be used to model the dynamics of copy, mutate, append, and extend events.

2.0 INTRODUCTION

Previous work on characterizing the dynamics of the spread of fake news were: (1) focused on social media (e.g., Twitter) and (2) do not consider the information content and rely instead on information such as tweet time and/or the number of retweets. The intents of this project were to (1) focus on news sources, not social media and (2) consider the language content of the news articles in addition to publication time to model how information evolves. One application of such an approach is to see if there is a quantifiable difference between fake news and real news.

We collected our own dataset as the existing large dataset often only contained the news headline, whereas we needed the raw text. We search the internet for news coverage of a particular event by keywords and then collect all the articles on that event (raw text, article title, metadata, etc.). We collected (English) articles from all the news outlets that appeared in the search ranging from major news outlets (e.g., CNN, BBC, FOX, etc.), gossip news outlets (e.g., TMZ, EONLINE, etc.), foreign news outlets that published English articles. A lot of effort was spent on methods to clean the collected dataset. We collected data on three major news events: (1) Jussie Smollett incident, (2) Ukraine International Airlines Flight 752, (3) COVID news in early 2020.

To study the information content of the article text, we used existing natural language processing (NLP) tools to extract semantic triples from articles collected each day. We organized the triples from articles collected each day as an RDF graph. Therefore, a corpus over time (e.g., news articles collected on different days) is described by a sequence of graphs.

We characterized the dynamics of the RDF graph by counting four types of features to describe how the structure of the cumulative RDF graph changes over time: Copy, Mutate, Append, and Extend. We modeled the dynamics using a discrete-time, multivariate Hawkes process.

Given the size of the collected dataset, it was not easy to discern 'fake' news from real news. Therefore, we divided the collected data on the Jussie Smollett incident into articles from major news outlets and celebrity news outlets. We showed that analyzing the evolution of RDF graphs shows both dynamic and information differences in the different news outlets.

We also showed that the discrete-time, multivariate Hawkes process can be used to model the dynamics; however, we did not investigate whether the Hawkes process can be used to predict future news dynamics (in the absence of breaking news development, which would be very hard to foresee).

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Data Collection

News articles are scrapped daily from the internet using the links provided by two services NewsAPI (https://newsapi.org) and Event Registry (https://eventregistry.org/). These services crawl the internet and curate HTML of articles based on an event query, which we provided. The results of both services are slightly different; therefore, we switched to Event Registry only. The cleaned dataset consists of three significant news events: (1) Jussie Smollett incident, (2) Ukraine International Airlines Flight 752, and (3) COVID news in early 2020.

Table 1. Datasets Collected

Dataset Name	Period of Collection	Num. News Outlets	Num. of Articles
Jussie Smollett	1/28/2019 – 3/14/2019	1,009	11, 340
Flight 752	1/2/2020 – 2/1/2020	2,136	33,063
COVID	1/8/2020 – 4/9/2020	10,449	2,838,324

3.2 Data Extraction and Cleaning

3.2.1 Data Extraction

We used newspaper3k (https://newspaper.readthedocs.io/en/latest/) to extract the following data from the HTML codes. In addition to the article text, we collected the following metadata from each article: URL, publication date, article title, authors, keywords (of the HTML page). An example below:

- **URL:** https://abc7.com/5823575/
- **Publication Date:** 2020-01-07 00:00:00
- **Article Title:** "US, Iran step back from the brink; President Trump opts for sanctions"
- **Authors:** []
- **Keywords:** [`strike', `trump', `sanctions', `step', `opts', `iran', `troops', `region', `forces', `missiles', `brink', `military', `american', `president']

We can see that not all fields in the metadata are available. For example, the author list is missing in this example, and the publication date defaulted to midnight of the day of publication.

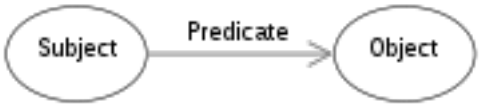
3.2.2 Text Cleaning

We do a series of simple cleaning of the extracted text by removing: (1) emojis/flags/other pictographs, (2) extra spaces, (3) parentheses and brackets.

3.3 RDF Graph Representation of Article Text

Natural Language Processing (NLP) algorithms are designed to extract information automatically from human-generated text. We used the ReVerb program (http://reverb.cs.washington.edu) to extract semantic triple (also called RDF triple) from each article text. A semantic triple extracts a **(subject; predicate; object)** expression from a sentence.

We can also view a semantic triple as a directed graph with the two end nodes representing the **subject** and **object** phrases as shown below:



Unlike classic Relation Extraction and Linking (REL) approaches, we **do not** link the extracted triples to any KB (knowledge base) such as DBpedia. There are several reasons for this. The primary reason is that we are not building an RDF graph for query purposes. Instead, we are interested in the properties and dynamics of the RDF graphs to characterize news articles over time. An additional reason is that we hoped to avoid the computational cost of linking the extracted entities to KB.

One disadvantage of not linking entities is that the extracted triples are very noisy. We encountered the following issues:

- Lots of pronouns (e.g., him', 'it', 'this', 'he', 'they', 'his', 'us', 'her', 'she', 'their', 'we', 'my', 'its', 'himself', 'them', 'that', 'which', 'i', 'you') in the article text.
- Many of the extracted triples have semantically identical/similar components. In the simple cases for example, ‘jussie smollett’ and ‘smollett’ are identified as two separate entities (i.e., two separate nodes or edges in the RDF graph). In more complex cases, it

can be very difficult to determine if two extracted entities contain the same semantic information.

To replace pronouns with the most likely referred to entity, we first use the Neuralcoref from the SpaCy library [1].

As it does not link all the pronouns, we add additional cleaning steps. Because these steps involve pairwise comparison between all the extracted entities, we cannot apply it to all the collected data for computational reasons. Typically, we choose a subset of triples (i.e., sub-RDF graph) and apply the following additional steps:

1) When two nodes (i.e., subject or object) have more than 50% words in common, we collapse them into a single node. Special care is considered on edges (i.e., predicate) for the word ‘not’.

2) We use word2vec() in Gensim [2] to project triples into vectors and evaluate their similarities. From experimentation, we decided that nodes/edges with similarity score > 0.8 should be collapsed into one node/edge.

For example, the following subject/object phrases 'antiblack hate crimes', 'some hate mail', 'real victims of hate crimes', 'an antigay and antiblack hate crime', 'the only hoax hate crime', 'some specializing in hate crimes', 'the seemingly sympathetic victim of a hate crime' were all reduced to the phrase ‘a possible hate crime’.

3.4 Fundamentals of Network Science to Analyze RDF graphs

Let G_i denote the RDF graph built from all the articles collected on the ith day pertaining to a specific event.

Therefore, G_1 is the RDF graph built from all the collected articles published on the first day. G_2 is the RDF graph built from all the collected articles published on the second day, etc.

As the number of articles (and consequently the number of triples) grows very quickly over time, the size of the RDF graphs increases rapidly as well.

It becomes very difficult to visualize the structures of the RDF graph. Therefore, we use tools from network science. Network science is the field of analyzing statistical properties of large graphs such as social networks, protein interaction graphs, and others [3]. One major downside of network science analysis is that the words associated with the semantic triples are not considered. However, the structure of the RDF graphs does reflect some semantic information in that edges are triples. Therefore, the connectivity of the edges reflects relationship between triples.

The second problem is that triples may have different Predicates but the same Subject and Object phrases. As a result, RDF graphs are multigraphs in that multiple edges can exist between two nodes. However, network science analysis is traditionally done on simple graphs (i.e., non-

multigraphs). Some of the network science characteristics we considered were Number of Cycles, Average Path Length, and Diameter as described below.

3.4.1 Number of Cycles

A cycle graph is a special graph structure where a set of vertices are connected in a closed chain. The triangle is the smallest possible cycle graph. The triangle structure is especially important in social networks due to triadic closure [4]. The triangle structure is also an important characteristic in network density and community detection. A dense graph has many triangles.

3.4.2 Average Path Length

The path length between any two nodes in a graph is the number of edges in the shortest path between the two nodes. In any given (undirected) graph with N nodes, there are N choose 2 possible pairs of nodes. The average of the path lengths between all these pairs of nodes is the average path length. A network that is 'well-connected' would have a short average path length. Social networks naturally evolve so that they have a short average path length.

3.4.3 Diameter

The diameter of a graph is the length of the longest shortest path between any two nodes. It is also a measure of connectivity in that a 'well-connected' network would have a small diameter.

The next section shows that RDF graphs have very different network characteristics from social networks. This informs us of the structure of RDF graphs and how they may change over time.

3.5 Modeling Dynamics of RDF Graphs Fundamentals of Hawkes Process for Dynamics Process

3.5.1 Dynamic Cumulative RDF

Let G_i denote the RDF graph built from all the articles collected on day *i* pertaining to a specific event. Therefore, G_1 is the RDF graph built from all the collected articles published on the first day. G_2 is the RDF graph built from all the collected articles published on the second day, etc. To analyze how the set of articles over multiple days, we can study the cumulative RDF by composing the RDF graph collected each day into a larger graph.

Let $G_1^T = G_1 \cup G_2 \cup \ldots \cup G_T$ denote the cumulative RDF built from daily RDF graphs. As the cumulative RDF graph may feature several disconnected subgraphs, we study the largest connected subgraph.

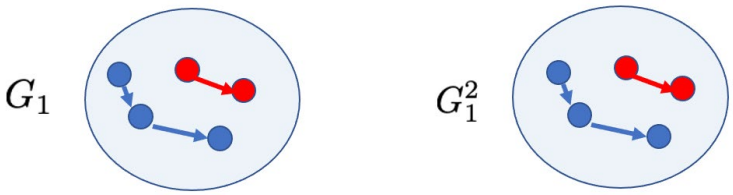
Ideally, we would like to model the dynamics of the cumulative RDF graph. However, modeling dynamical graphs is very difficult due to the high dimensionality of the structure. Therefore, we selected four specific features of the cumulative RDF graph to model as multivariate time-series.

The four events types are chosen based on how we think the cumulative RDF graph might grow over time.

3.5.2 Event Types

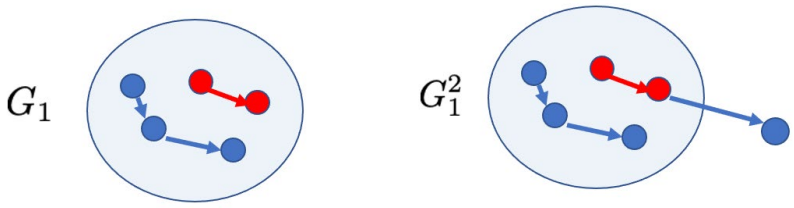
Consider a (subject; predicate; object) triple in an initial RDF graph G_1. This initial graph can change in 4 different ways to the cumulative RDF graph $G_1^2 = G_1 \cup G_2$ (i.e., the cumulative RDF graph that includes triples generated from articles collected on day 1 and day 2).

1. **Copy:** the new triple is identical to (subject; predicate; object)

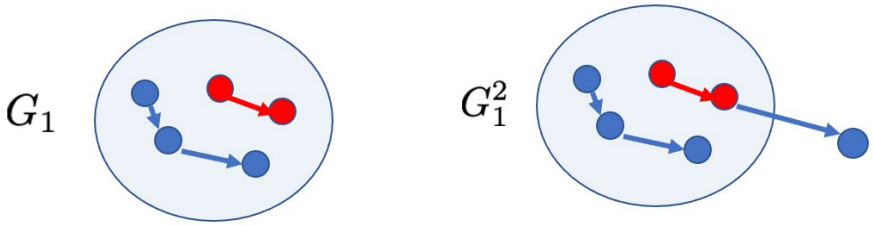


2. **Mutate:** the new triple has the same subject and object but different predicate

3. **Append:** new triple either appends to the head or tail node of (subject; predicate; object)



4. **Extend:** a new predicate connects two disconnected triples



Copy and mutate do not change the structure of the cumulative RDF graph (if we discount multiple edges), whereas append and extend result in a structural change of the graph.

We can model the dynamics of the cumulative RDF graph by looking at the time series corresponding to the number of times each particular event occurs on the nth day. For example, if the number of copy events is 3 on day 4, this means that the RDF graph on day 4 has 3 triples that are identical to some triples in the cumulative RDF graph G_1^3.

3.6 Hawkes Process Models

We chose to use the Hawkes process to model the dynamics of copy, mutate, append, and extend events for the following reasons: 1) Hawkes process is a self-exciting process. The probability of new event occurrences increases the probability of future event occurrences. As a result, Hawkes process has been used to model the dynamics of earthquake occurrences, social media posts [5] [6]; 2) multivariate Hawkes process can be used to model the dynamics of multiple (interdependent) time series [7]; 3) Hawkes process can be modified into a discrete-time process [8].

The disadvantage of the (multivariate) Hawkes process is that the parameters of the Hawkes can be hard to learn. The maximum likelihood estimator (MLE) is a standard method for learning the parameters. However, the likelihood function of the Hawkes process is not convex. Furthermore, we need a sufficient number of observations in order to estimate the parameters; this may not be satisfied in real-world datasets.

3.6.1 Continuous-time Hawkes Process

Consider N(t), t > 0, a counting process (a stochastic process that counts the total number of event arrivals/occurrences from time 0 to time *t*) with history $\mathcal{H}_t$ that satisfies:

$$P(N(t+h) - N(t) = m \mid \mathcal{H}_t) = \begin{cases} \lambda^*(t) + o(h), & m = 1 \\ o(h), & m > 1 \\ 1 - \lambda^*(t) + o(h), & m = 0 \end{cases} \quad (1)$$

A self-exciting process is a special type of point process where the event history of the process can increase future probability of event occurrence. The most often-used self-exciting process is the Hawkes process. The Hawkes process is a counting process, N(t), that models the number of arrival of some event in a time interval. It has been used to model the number of earthquakes, trade orders, social media posts, etc. [5]. Hawkes process is a non-Markovian extension of the Poisson process. Whereas the Poisson process is parameterized by a constant intensity (also called rate), the Hawkes process is parameterized by a conditional intensity function, $\lambda^*(t)$.

The parameter $\lambda^*(t)$ is a conditional intensity function because it depends on the past history of the process up to (but not including) time *t* denoted as $\mathcal{H}_t$. The past history of the process is the time of all the event arrivals up to time *t,* $\mathcal{H}_t = \{t_i : t_i < t\}$. One way to interpret $\lambda^*(t)$ is that it is the expected number of observed events in a small-time interval of length *dt* given $\mathcal{H}_t$.

$$\lambda^*(t) = \frac{E[dN(t) \mid \mathcal{H}_t]}{dt}, \mathcal{H}_t = \{t_i : t_i < t\} \quad (2)$$

The Hawkes process is a continuous-time stochastic process. This means that multiple events **cannot** occur simultaneously; this assumption is reasonable if observations can be made very quickly but may not be satisfied in real-world data due to the low sampling rate. The Hawkes process assumes that the conditional intensity function is of the general form:

$$\lambda^*(t) = \lambda + \sum_{t_i<t} \phi(t - t_i) \quad (3)$$

where $\{t_1, t_2, \ldots t_k\}$ denotes the observed sequence of event arrival times up to time t, $\lambda > 0$ is the background intensity. $\phi()$ is the excitation function (also known as the memory kernel function), which can take on several forms. The two popular forms of the kernel function are:

1) The exponential kernel:

$$\phi(t - t_i) = \alpha e^{-\beta(t-t_i)}, \alpha, \beta > 0 \quad (4)$$

2) The power-law kernel:

$$\phi(t - t_i) = \frac{k}{c+(t-t_i)^p}, c, k, p > 0 \quad (5)$$

3.6.2 Discrete-time Hawkes Process

The classic Hawkes process is a continuous-time model. This means that events cannot occur simultaneously; this is a reasonable assumption for a continuous-time system. For example, the first arrival is at time t=1, and the second arrival is at time t=1.01. However, real-world data may be collected only at discrete time instances. For example, the data we are working with collected per day. It is very likely then that simultaneous events arrive in the same day. Therefore, we must work with a discrete-time variation of the Hawkes process, introduced in reference [8].

Whereas the classic continuous-time Hawkes process is a counting process N(t), which models the cumulative number of event occurrences up to time t, it is easier to interpret the discrete-time Hawkes process from the perspective of a stochastic process Y(t), which models the number of event occurrences on day t. Note then that:

$$Y(t) = N(t) - N(t-1) \quad (6)$$

The interpretation of the **conditional intensity function,** $\lambda^*(t)$ (in univariate case) and $\Lambda^*(t)$ (in multivariate case), and the **history,** $\mathcal{H}_t$, are slightly different for discrete-time Hawkes process.

The history of the discrete-time process, $\mathcal{H}_t$, is not a collection of observed arrival times (since multiple events can arrive at the same time) but the number of observed arrivals each day from day 0 to day t-1: $\mathcal{H}_{t-1} = \{Y(s) = y(s) | s \leq t-1\}$. y_s represents the observed number of events on day s; it is the observed instance of the random variable Y(s).

We can interpret the conditional intensity function, $\lambda^*(t)$, of the discrete-time process as the expected number of events that occurs on day *t* given the history $\mathcal{H}_{t-1}$.

$$\lambda^*(t) = E[Y(t)|\mathcal{H}_{t-1}], \mathcal{H}_{t-1} = \{Y(s) = y(s) | s \leq t-1\} \quad (7)$$

The discrete-time Hawkes process assumes that the conditional intensity function is of the general form:

$$\lambda^*(t) = \lambda + \alpha \sum_{t_i < t} y(t_i)\phi(t - t_i) \quad (8)$$

Where λ is the base conditional intensity, α is a scaling factor, y_s is the observed number of events on day s, and $\phi(t - t_i)$ is the excitation function. A common choice of excitation function for the discrete-time Hawkes process is:

$$\phi(t - t_i) = \beta(1 - \beta)^{t - t_i} \quad (9)$$

Or

$$\phi(t - t_i) = e^{-\beta(t - t_i)} \quad (10)$$

In the discrete-time Hawkes process, Y(t), the number of event occurrences on day t, is a Poisson random variable with mean $\lambda^*(t)$, that is:

$$P(Y(t) = y) = \frac{(\lambda^*(t))^y e^{-\lambda^*(t)}}{y!} \quad (11)$$

3.6.3 Multivariate Hawkes Process

Classic Hawkes' process models the arrival/occurrence of one event type (e.g., earthquake). A multivariate Hawkes process is a d-dimensional counting process $N(t) = \{N_1(t), N_2(t), \ldots N_d(t)\}$ and models the arrival of *d* different event types. The assumption is that the arrival of one type of event depends not only on its own history but also the history of the other types of events. Each type of event is characterized by its own conditional intensity function: $\lambda_1^*(t), \lambda_2^*(t), \ldots \lambda_d^*(t)$ so that:

$$\lambda_i^*(t) = \lambda_i + \sum_{j=1}^{d} \sum_{t_i < t} y_i(t_i)\alpha_{i,j}\phi_i(t - t_i) \quad (12)$$

3.6.3.1 Parameter Estimation of Discrete-time Hawkes Process

To use the Hawkes process, we have to estimate all the parameters of the conditional intensity function (i.e., $\theta = \{\lambda, \alpha, \beta\}$) from observations. One approach to estimate these parameters is the maximum likelihood estimator (MLE) because the Hawkes process has a closed-form likelihood function. However, the likelihood function is not convex and can be difficult to optimize.

As the likelihood function of the continuous-time Hawkes process can be theoretical challenging to justify (due to infinitesimal length time intervals) and we only used discrete-time Hawkes process because of the nature of our data, we will only discuss parameter estimation of discrete-time, univariate Hawkes process and parameter estimation of discrete-time, multivariate Hawkes process.

Given one sequence of observations **y** = {y(1), y(2), y(3) … y(N)}, the likelihood function of the univariate discrete-time Hawkes process is:

$$L(\mathbf{y}|\theta) = \mathrm{P}(\mathrm{Y}(1) = \mathrm{y}(1), \ldots \mathrm{Y}(\mathrm{N}) = \mathrm{y}(\mathrm{N})|\theta)$$
$$= P(Y(1) = y(1) \mid \lambda^*(1))P(Y(2)\ y(2)|\ \lambda^*(2))\ldots P(Y(N) = y(N)|\lambda^*(N))$$
$$= \prod_{n=1}^{N} P(Y(n) = y(n)|\lambda^*(n)) \quad (13)$$

As the increments are conditionally independent given the condition intensity function, $\lambda^*(t)$, which depends on $\theta = \{\lambda, \alpha, \beta\}$, the log-likelihood is then:

$$logL(\mathbf{y}|\theta) = \sum_{n=1}^{N} y(n)log(\lambda^*(n)) - \lambda^*(t) - log(y(n)!) \quad (14)$$

The maximum likelihood estimator for θ is:

$$\breve{\theta} = arg \max_{\theta} \sum_{n=1}^{N} y(n)log(\lambda^*(n)) - \lambda^*(t) \quad (15)$$

We do not need the factorial expression of y(n) because it is not a function of θ.

3.6.3.2 Parameter Estimation of Discrete-time Multivariate Hawkes Process

Observations of multivariate Hawkes process has *d* different sequences:

$$\boldsymbol{y}_1 = \{y_1(1), y_1(2), y_1(3), \ldots y_1(N)\},$$
$$\boldsymbol{y}_2 = \{y_2(1), y_2(2), y_2(3), \ldots y_2(N)\}$$
$$\ldots$$
$$\boldsymbol{y}_d = \{y_d(1), y_d(2), y_d(3), \ldots y_d(N)\}$$

Each type of event is characterized by its own conditional intensity function:

$$\lambda_1^*(t), \lambda_2^*(t), \ldots \lambda_d^*(t)$$
$$\theta = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{21}, \alpha_{22}, \ldots, \alpha_{43}, \alpha_{44}, \beta_1, \beta_2, \beta_3, \beta_4\}$$

As the increments are conditionally independent given the condition intensity function, which depends on θ, the log-likelihood is then:

$$logL(\boldsymbol{y}_1, \ldots \boldsymbol{y}_d|\theta) = \sum_{i=1}^{d} \sum_{n=1}^{N} y_i(n)log({\lambda_i}^*(n)) - {\lambda_i}^*(t) \quad (17)$$

4.0 RESULTS AND DISCUSSION

One simple way to look at the dynamic of the event is to look at the plot of the number of articles published/collected per day. The three collected dataset has different behavior due to how the events evolved.

1) **Jussie Smollett** dataset has two peaks in the number of articles published for the event; interestingly, more articles were published during the second peak than the initial event.

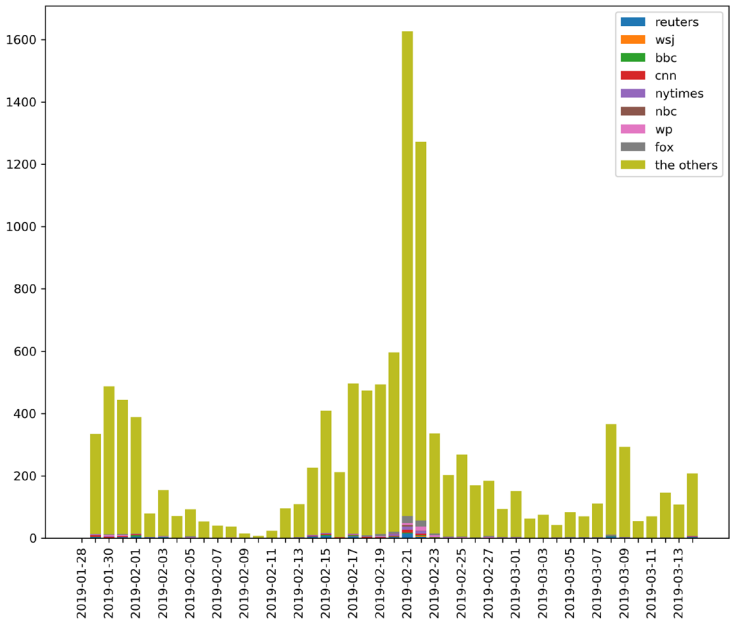


Figure 1: Number of articles published per day (Jussie Smollett)

2) **Flight 752** dataset only has one peak; presumably, most news events would share this short of dynamics where coverage increases and then gradually decreases.

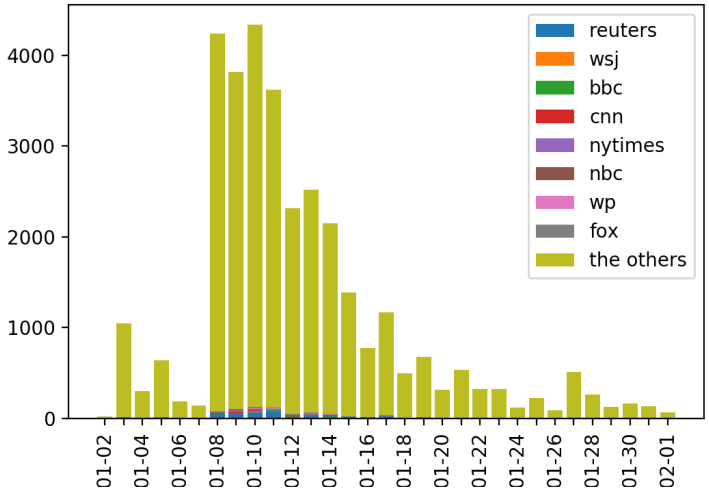


Figure 2: Number of articles published per day (Flight 752)

3) Our collected **Covid19** news articles covered only the first 4 months in 2020 because the number of articles published each day quickly overwhelmed then collection effort. Additionally, the keyword 'Covid19' may be too broad since it could refer to a number of different news coverage.

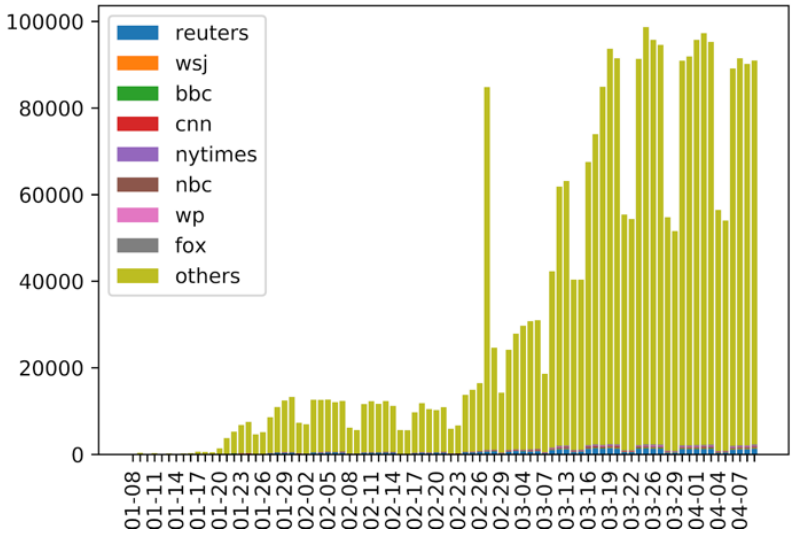


Figure 3: Number of articles published per day (COVID19)

Note: due to the size of the dataset, we spend a significant amount of time cleaning (see Section 3.2.2) and working with the Jussie Smollett dataset. The other two datasets have yet to be cleaned and analyzed.

We conducted three analyses with the Jussie Smollett dataset:

(1) Our first experiment uses network science statistics detailed in Section 3.4 to gain a sense of the structure of the RDF graph built from the set of articles collected each day.

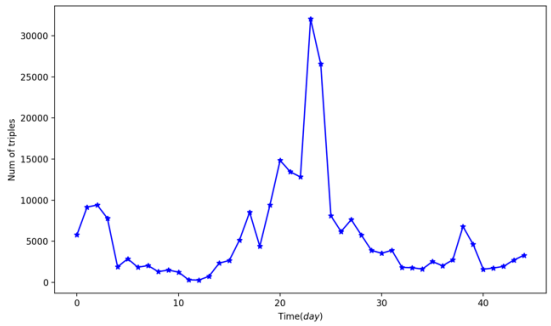
(2) Our second experiment is to see if RDF graphs give more information than article counts. As the RDF graphs are too large and complicated to be directly visualized, we used the dynamics of the four different event types: (i.e., Copy, Mutate, Append, and Extend). **Our hypothesis is that articles published by major news outlets and celebrity gossip news outlets would exhibit a difference in semantic information and in article counts.** In comparison, we created a control set where we randomly generated two sets of articles. We will see that the experiment involving **major news outlets** and **celebrity news outlets** show less correlation between article count and RDF events dynamics.

(3) Our third experiment is to model the dynamics of the four different event types (i.e., Copy, Mutate, Append, and Extend) using a discrete-time multivariate Hawkes process model (see Section 3.6. The reason that we chose the Hawkes process to model the time series is because it is a self-exciting process and has been used to model the dynamics of social media post. It can easily be extended to a multivariate process where we can model all 4 event type time series and how they may affect one another.

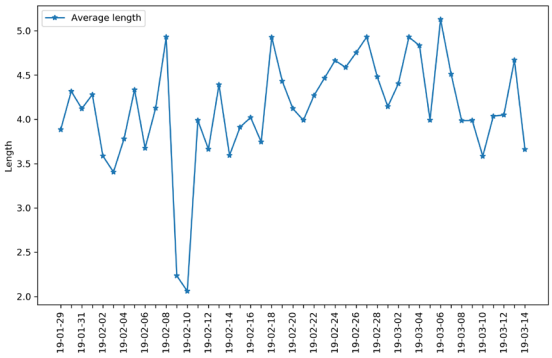
To use the data that we collected, we had to use the discrete-time variation of the Hawkes process as we only have the date of publication for all the articles.

Furthermore, as the dynamic of the process is nonstationary, we obtained the best performance when we model a very short time window. In the case of unusual development, such as the news of Jussie Smollett's indictment in Feb. 2019, a statistical method cannot predict the occurrence of such an event.

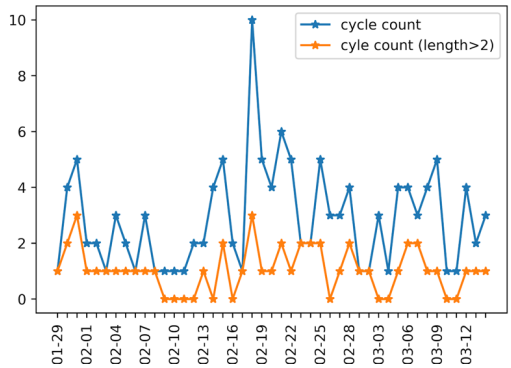
4.1 Network Science Analysis of RDF Graph Structures



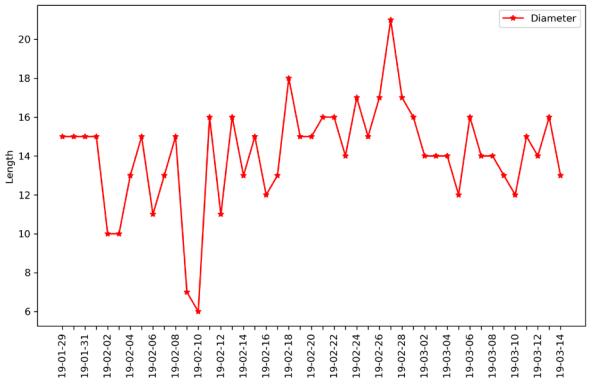
(a) Number of triples (i.e., edges) extracted each day



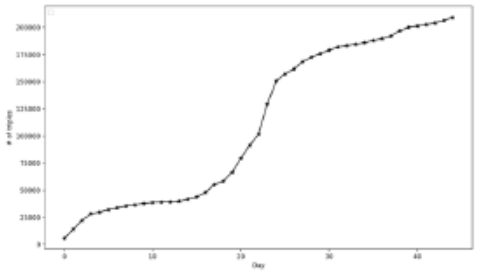
(b) Average path length of RDF graph (largest connected component) each day



(c) Number of Cycles in RDF graph (largest connected component) each day



(d) Diameter of RDF graph (largest connected component) each day



(e) Cumulative number of triples over the entire collection period

Figure 4: Network Statistics over RDF graphs

To begin our analysis of RDF graphs built from triples extracted from articles collected each day, we looked at some basic network statistics as described in Section 3.4.

We can see that RDF graphs have a low number of cycles (unlike social network communities), relatively short diameter, and average path length. Roughly, we expect the RDF to have a star-hub structure where a few nodes have many neighbors, which are unconnected between themselves.

This is because a few subject/object phrases appeared throughout coverage (e.g., 'Smollett' and 'police').

4.2 Dynamic Difference between Major News Outlets vs. Gossip News Outlets

4.2.1 Random Set of Articles (Control)

We first conducted a preliminary experiment to show that two sets of articles from randomly chosen news outlets exhibit similar dynamics as article counts (as expected). We built **set1** with 2,042 articles from randomly selected news outlets and **set2** with 2,033 articles from randomly selected news outlets. To make the comparison between different quantities easier, we rescaled all the values from 0 to 100. We note in each figure caption if the plotted values are rescaled from the original values or not.

4.2.2 Dynamics of Article Counts

A good deal of dynamic information can be gained by looking at the number of articles published over time; we were careful to pick set1 and set2 so they have approximately the same number of articles. We start the plot in day 2 because the RDF graphs keep track of changes starting in day 2.

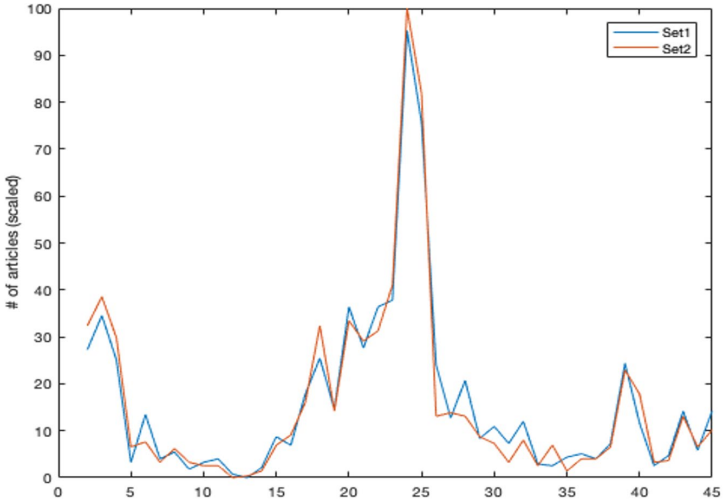
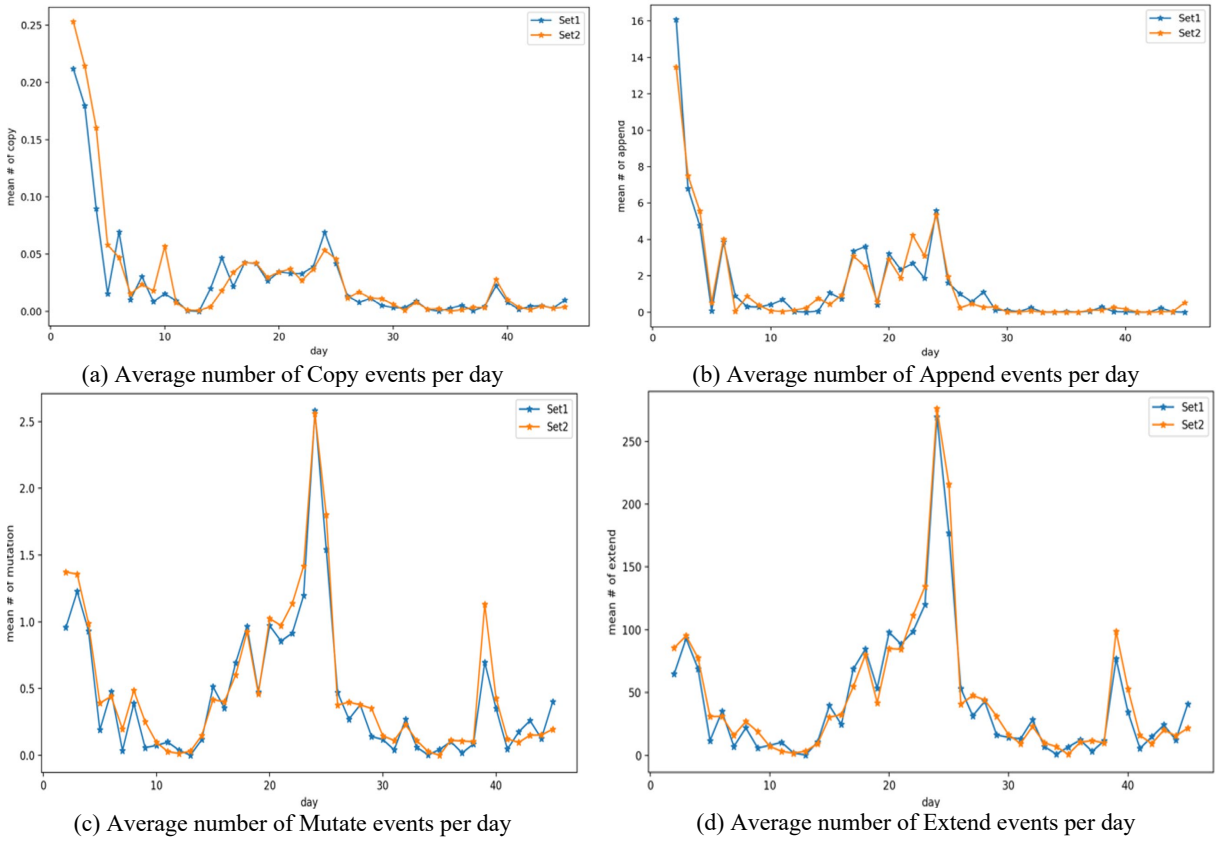


Figure 5: Number of articles published per day (scaled from 0 to 100)

4.2.3 Dynamic of RDF Graph Changes



(a) Average number of Copy events per day

(b) Average number of Append events per day

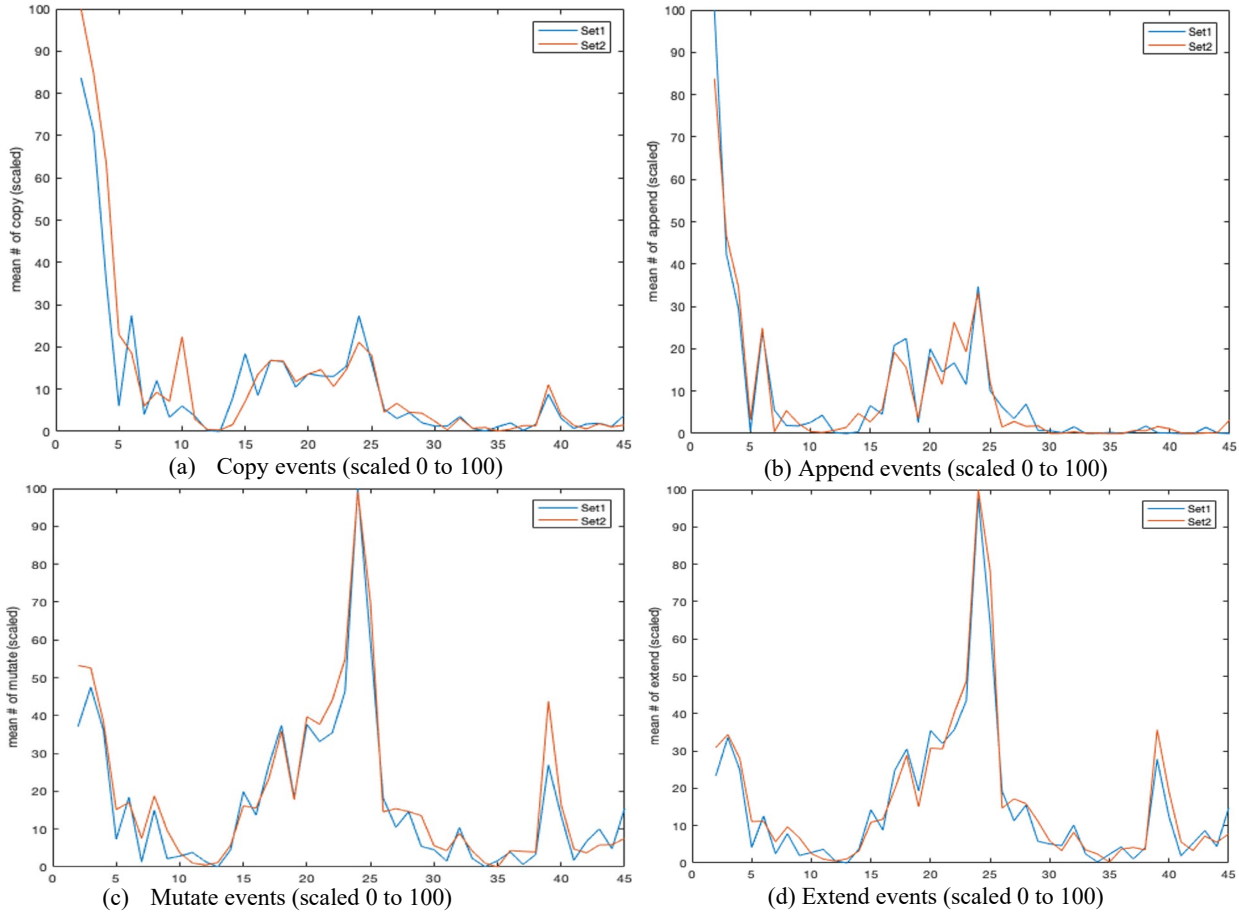
(c) Average number of Mutate events per day

(d) Average number of Extend events per day

Figure 6: Average (a) Copy, (b) Append, (c) Mutate, (d) Extend events per day for set1 and set2.

We generated the RDF graphs for each set using the method discussed in Section 3.3 and cleaned the data according to the procedures discussed in Section 3.2.2. We then characterized the dynamics of RDF graphs over time by noting the number of times the four different events (i.e.,

Copy, Mutate, Append, and Extend) occur as defined in Section 3.5.2. Figure 6 shows the plot for average number of (a) Copy, (b) Append, (c) Mutate, and (d) Extend events per day starting on day 2. Each triple in G_1 generates its own set of occurrences of different events. The average number of the different types of events are very similar for Set1 and Set2 as expected. The two set of values in the Figure 6 have very different range; therefore, we felt it was important to show the rescaled range of all the values from 0 to 100 as shown in Figure 7.



(a) Copy events (scaled 0 to 100)

(b) Append events (scaled 0 to 100)

(c) Mutate events (scaled 0 to 100)

(d) Extend events (scaled 0 to 100)

Figure 7: Average number of extend events per day for set1 and set2 scaled from 0 to 100.

We then compute the absolute value difference between article count and the average number of different RDF event types for set1 and set2. As the curves are rather jagged, it is hard to detect trends. A 5-day moving average is used to smooth out the curves. This means that each value is the average of the values in a 5-day window.

We can see that the curves for RDF event types are **highly correlated** with the article count. Because set1 and set2 are randomly generated, we do not expect much semantic differences between the articles in the two sets.

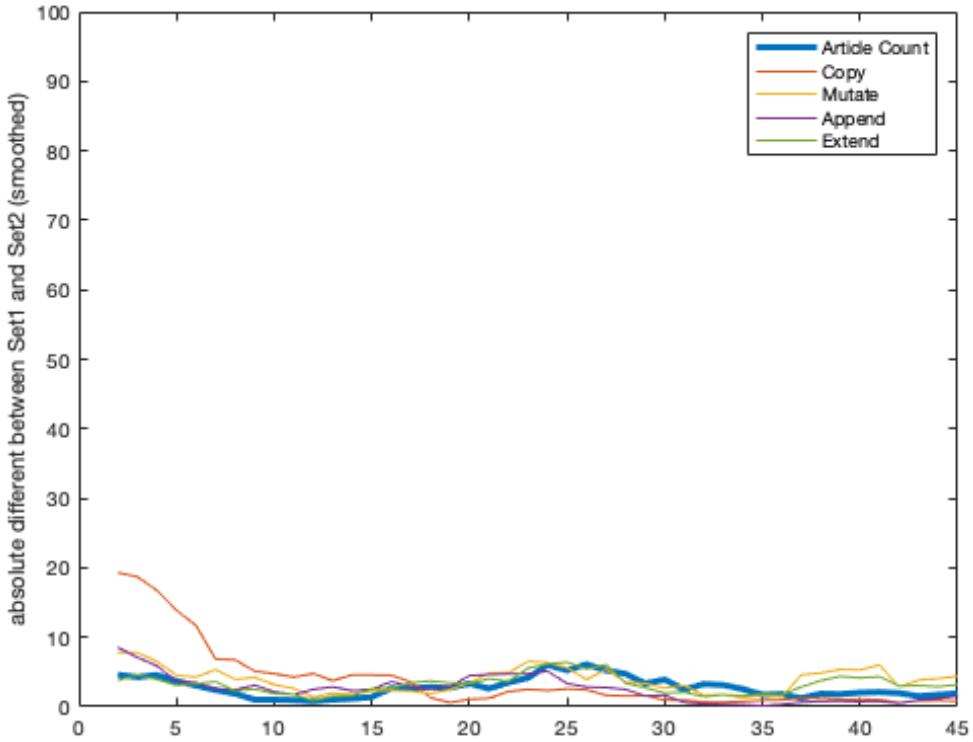


Figure 8: Absolute difference between Set1 and Set2 (smoothed by 5-day moving average)

4.2.4 Major News Outlets vs. Gossip News Outlets

Instead of randomly picking sets of articles, we selected articles based on the publication outlets. As in the previous case, we rescaled the values to between 0 and 100.

We considered the following news sources as **major news outlets**. We obtained this list from https://blog.feedspot.com/usa_news_websites/: nytimes, huffpost, foxnews, usatoday, politico, yahoo, npr, latimes, breitbart, nypost, abcnews, nbcnews, bbc, cnn, reuters.

We considered the following news sources as **celebrity gossip news outlets**. We obtained this list from https://blog.feedspot.com/celebrity_gossip_blogs/: eonline,bet, tmz,people, hollywoodlife, mtonews, theshaderoom, filmfare, usmagazine, deadline, vanityfair, zimbio, realityblurb, pinkvilla, perezhilton, hellomagazine, justjared, toofab, bollywoodlife, stylecaster, thejasminebrand, ohnotheydidnt.livejournal, bossip, theybf, celebitchy, sandrarose, etcanada, radaronline, thehollywoodgossip, intouchweekly, socialitelife, missmalini, jimmystarsworld, lifeandstylemag, extratv, gofugyourself, closerweekly, laineygossip, terezowens, allaboutthetea, bckonline, chaospin, gossipcop, x17online, latina, soapoperadigest, chano8, karencivil, topbuzztrends, malecelebnews, privilegeamoah, gossipbucket, wesmirch, entertainmentwise, filmistory, bollywoodjournalist.

There were 699 articles from major news outlets and 685 articles from celebrity gossip news outlets.

4.2.4.1 Dynamics of Article Counts

We can see that there are differences in the article count between major news outlet articles and celebrity gossip news outlet articles, especially around the time of Jussie Smollett's indictment (second peak).

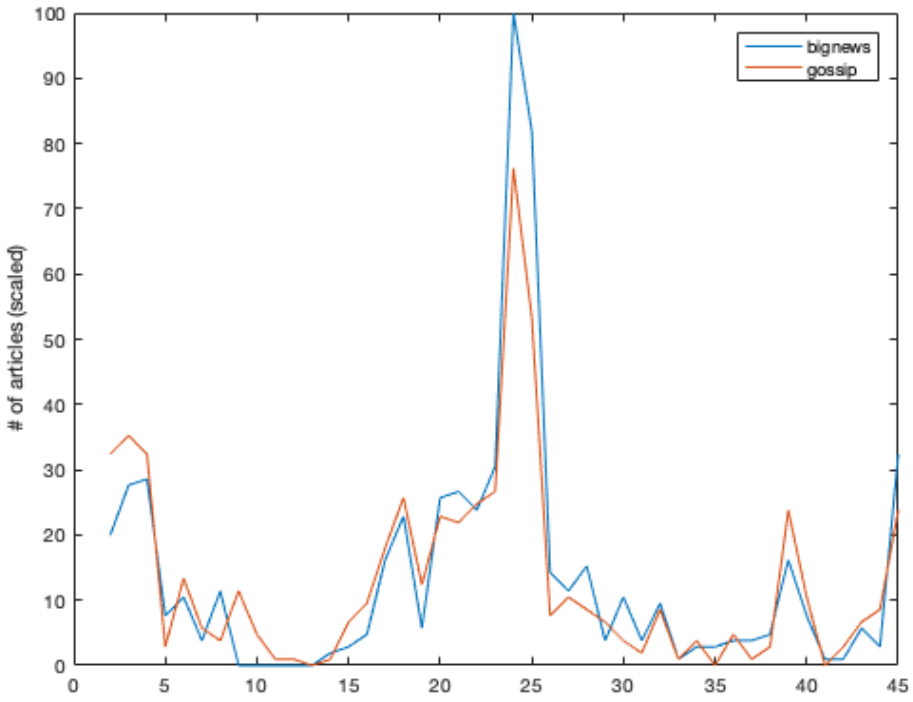
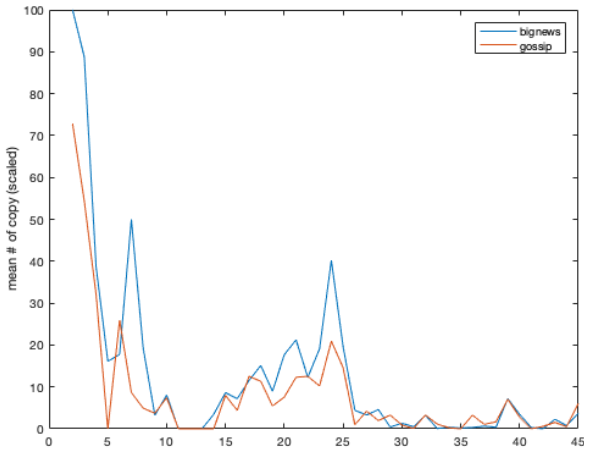


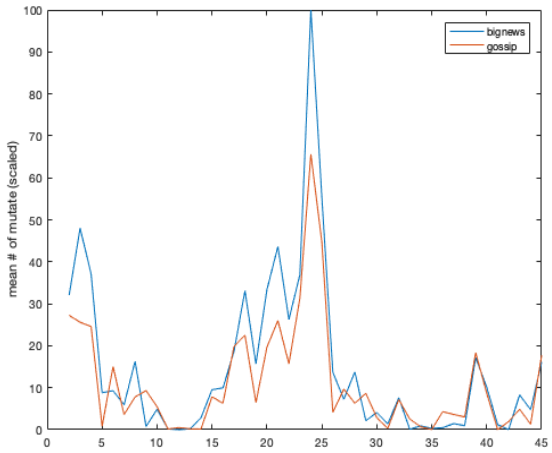
Figure 9: Number of articles per day (scaled from 0 to 100)

4.2.4.2 Dynamic of RDF Graph Changes

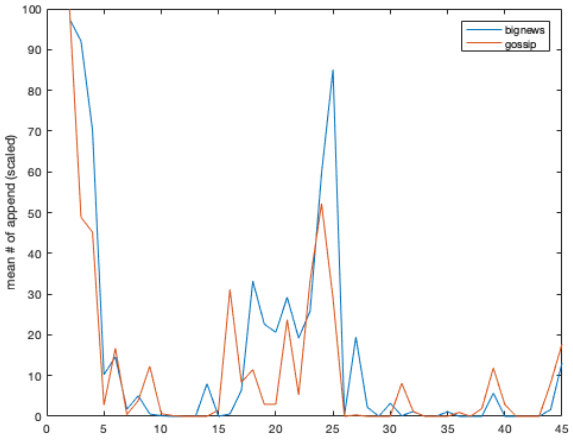
We generated the RDF graphs for each set using the method discussed in Section 3.3 and cleaned the data according to the procedures discussed in Section 3.2.2. We then characterized the dynamics of RDF graphs over time by noting the number of times the four different events (i.e., Copy, Mutate, Append, and Extend) occur as defined in Section 3.5.2. Figure 10 shows the scaled average of the number of Copy, Mutate, Append, and Extend events per day starting on day 2.



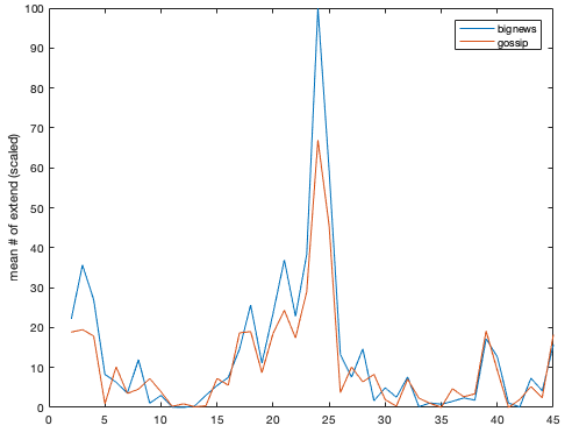
(a) Average number of Copy events per day (scaled from 0 to 100)



(b) Average number of Mutate events per day (scaled from 0 to 100)



(c) Average number of Append events per day (scaled from 0 to 100)



(d) Average number of Extend events per day (scaled from 0 to 100)

Figure 10: Average number of (a) Copy, (b) Append, (c) Mutate, and (d) Extend events per day between major news and gossip articles.

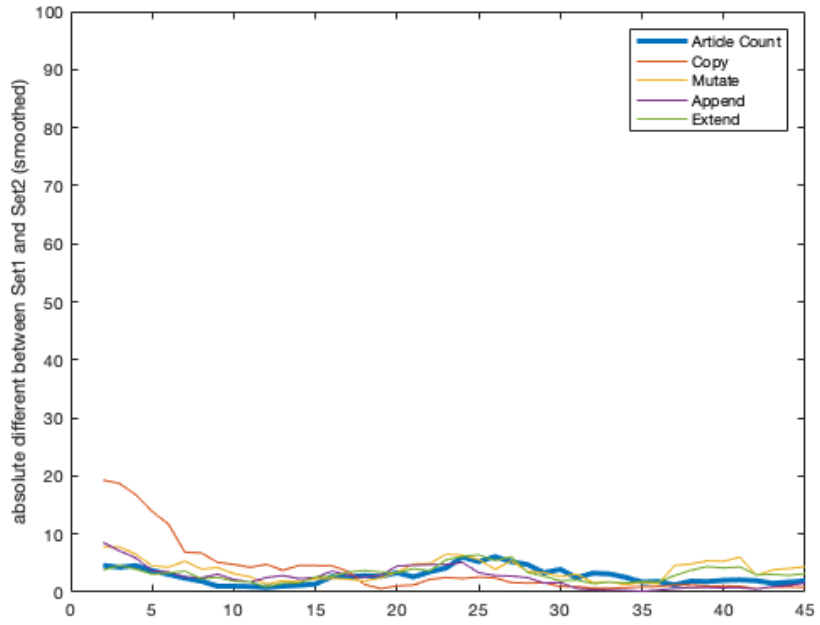
4.2.5 Random Corpus set 1 and 2 vs. Major News and Celebrity Gossip News

We can compare the absolute valued difference between article counts and the different RDF events between major news outlets and celebrity gossip news outlets (5-day moving average used to smooth out the difference curves).

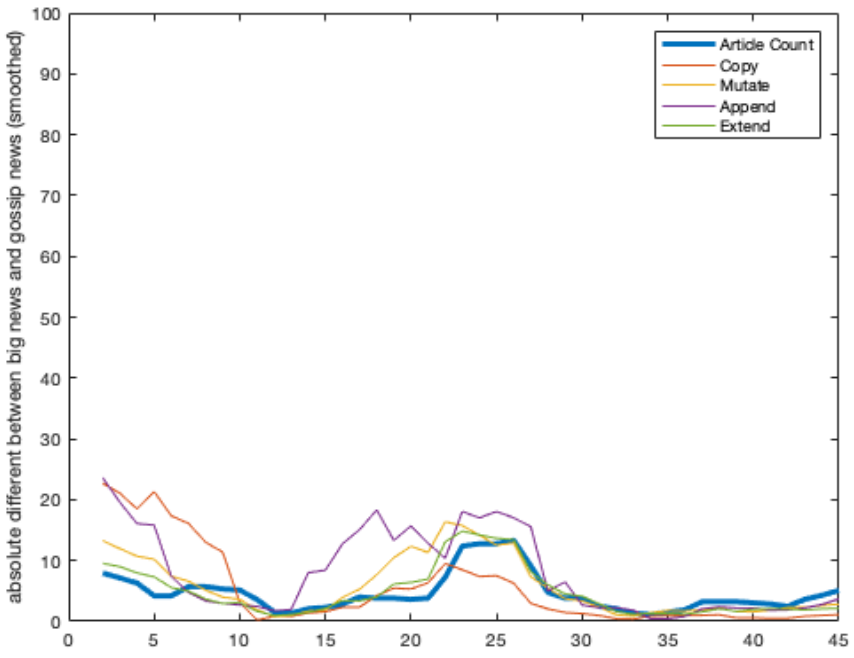
For the difference between major news outlets and celebrity news outlets, we can see that the curves for RDF event types are **less correlated** with article count than for set1/set2.

We expect more semantic differences between how major news outlets might have reported the story than celebrity gossip news outlets. In contrast, we don't expect articles from randomly selected outlets to have much of a difference. This shows that RDF events are picking up information in addition to the dynamics of article counts.

In particular, the append event dynamic is very different between major news outlets articles and celebrity gossip news outlet articles.



(a) Absolute difference between Set1 and Set2



(b) Absolute difference between major news outlets and gossip news outlets

Figure 11: Absolute difference between Set1 and Set2 vs. Major News Outlets and Gossip News (each smoothed by 5-day moving average)

4.3 Discrete-time, Multivariate Hawkes Process for Modeling and Predicting RDF Graph Dynamics

One major factor in building statistical models like the Hawkes process is to have sufficient data so that the parameter estimators can have low variance. This requirement is even more critical for the multivariate Hawkes process since we have to learn the pairwise parameters of how the time series dynamics affect the others.

The data we gathered are best modeled using discrete-time Hawkes process since the articles are published per day. In so far as we know, there is the first-time application of discrete-time, multivariate Hawkes process.

4.3.1 Learning the Parameters

Let *M* denote the number of triples from the initial point (i.e., the RDF graph built from triples collected on day 1, G_1). We treat the *M* triples are *M* independent sequences. Let *m* denotes a triple from G_1. On day 2, we count the number of events that the news triples extracted on day 2 in relation to *m*. For example, the observation of day 2 may be [1, 0 , 0, 2]; we know that there was 1 copy event to *m*, 0 mutate and append events, and 2 extend events to *m.*

Let m_2 denote this new graph structure. On day 3, we count the number of events that the news triples extracted on day in relation to m_2. For example, the observations of day 3 may be [1, 1, 0, 0]; we know that there was 1 copy event to m_2, 1 mutate event to m_2, and 0 append and extend events.

It is possible (and happens frequently) that the observation for a particular day will be [0,0,0,0]. As a result, even though the number of triples is very high each day, we may not have enough data for each event type. One way of preventing this from happening is to ensure that *M* (i.e., total number of initial triples) is sufficiently large.

Give a sequence of observations, we solved for the maximum likelihood estimator of the discrete-time, multivariate Hawkes process by maximizing the likelihood function using (what function?). As this is a nonconvex function, an efficient estimator remains an open research question. In our model we have the following parameters to estimate:

$$\theta = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{21}, \alpha_{22}, \ldots, \alpha_{43}, \alpha_{44}, \beta_1, \beta_2, \beta_3, \beta_4\}$$

As the likelihood function is nonconvex, the initial guess for the parameter values has a large impact on the result. **Furthermore, to simplify the maximization task, we assumed that $\boldsymbol{\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta = \sqrt{2}}$ is known.** Therefore, the only parameters we need to estimate are λ and α values (20 unknowns).

4.3.1.1 Dealing with Nonstationarity

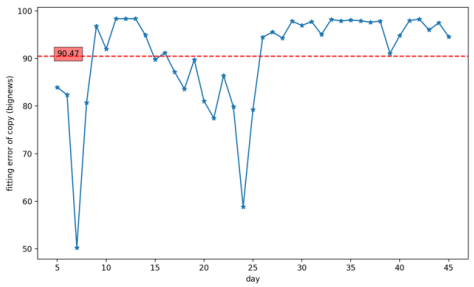
We can see from the plots of the average number of copy, mutate, append, and extend events in Section 4.2 that the values can change dramatically over time due to changing events. Therefore, we do not expect a Hawkes process model to be suitable for modeling the dynamics from day 1 until the end of the news cycle (for the Jussie Smollett story, that is around 40 days). As a result, we only consider using a Hawkes process to model the dynamics in a short time window. From experiments, we determined that a window of 10 days of change works best.

We collected observations of how the RDF graph changes from:

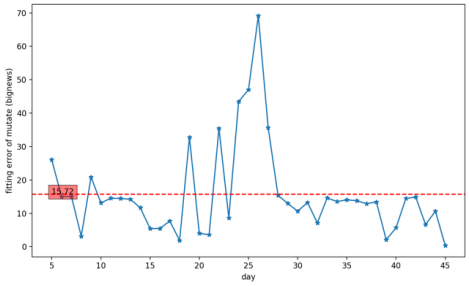
Day 1 to Day 10 (initial set of triples is from RDF graph G_1)
Day 7 to Day 16 (initial set of triples is from cumulative RDF graph G_1^7)
Day 13 to Day 22 (initial set of triples is from cumulative RDF graph G_1^{13})
…

4.3.2 Model Performance

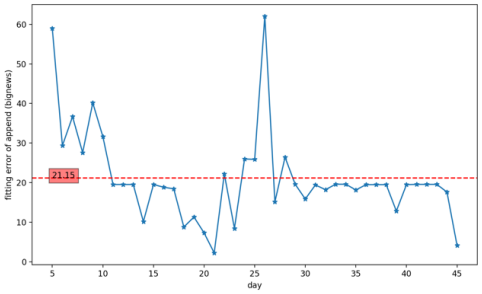
To compare the performance of using discrete-time, multivariate Hawkes process to model the dynamics of copy, mutate, append, and extend events, we compare the observed mean number of events for each type on day i, $y_1(i), y_2(i), y_3(i), y_4(i)$, with the estimated mean number of events for each type, $\check{y}_1(i), \check{y}_2(i), \check{y}_3(i), \check{y}_4(i)$. In order for the comparison to be fair, we scale the values to between 0 and 100. The estimated values are computed using the parameters θ learned for each window of observations from $y_1(i), y_2(i), y_3(i), y_4(i)$.



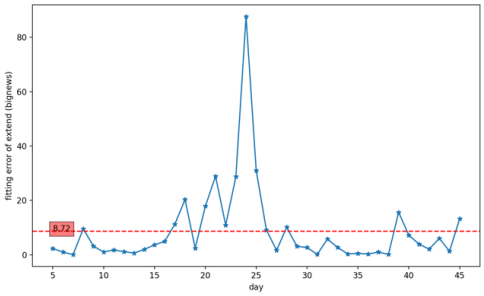
(a) Error for Copy events



(b) Error for Mutate events



(c) Error for Append events



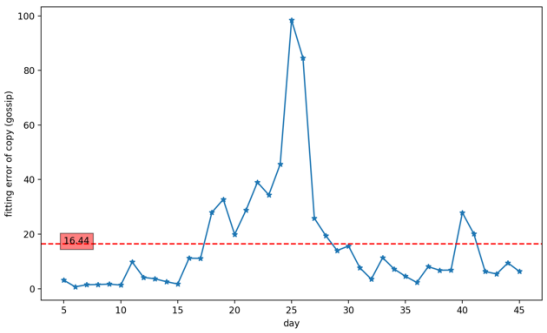
(d) Error for Extend events

Figure 12: Error between observed and estimated mean number of (a) Copy, (b) Mutate, (c) Append, and (d) Extend events of articles from major news outlets (each scaled from 0 to 100)

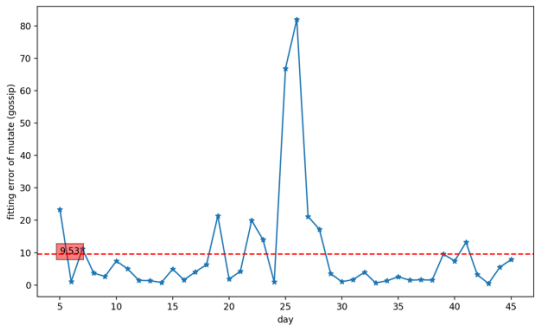
The error is defined as:

$$error = |y_d(i) - \check{y}_d(i)| \tag{18}$$

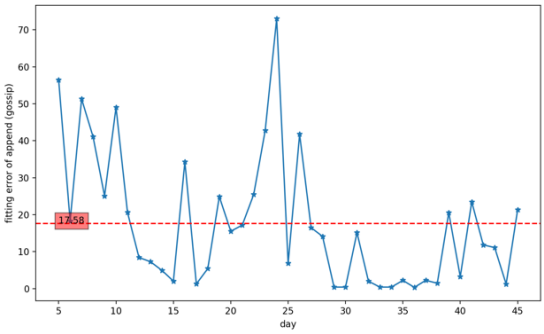
We considered the set of triples from major news outlets (Figure 12) and celebrity gossip news outlets (Figure 13) as dealing with the entire dataset makes parameter estimation very time consuming. The red horizontal dashed line shows the average error.



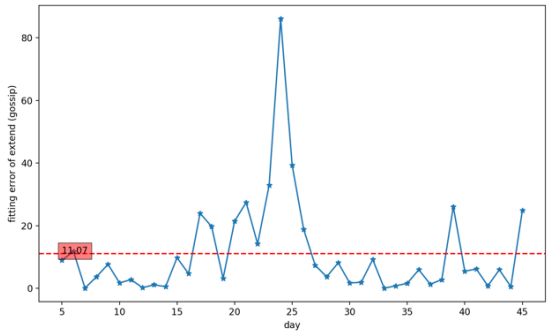
(a) Error for Copy events



(b) Error for Mutate events



(c) Error for Append events



(d) Error for Extend events

Figure 13: Error between observed and estimated mean number of (a) Copy, (b) Mutate, (c) Append, and (d) Extend events of articles from celebrity gossip news outlets (each scaled from 0 to 100)

We see that the performance of Hawkes process modeling is better for celebrity gossip news outlets than for major news outlets. In particular, the average error is very large for copy events in major news outlets. One possible reason is the lack of observations.

We also see that the largest divergence between observation and estimation is around the Feb 20-25, which is when the news of Jussie Smollett's indictment unexpectedly broke out. On these days, the estimated average value of event occurrences is smaller than the observed average value of event occurrences.

5. CONCLUSIONS

The goal of this project was to combine NLP analysis with dynamics modeling to characterize how news change over time (potentially) for detecting fake news. This is a novel line of research as most NLP problems are static data analysis problems, whereas dynamics modeling problems are numerical in nature.

The project had to overcome several challenges:

1. We collected a new dataset where a single news event was followed continuously over time from many news outlets. Most NLP datasets are collections of text without time stamps. Furthermore, we collected the entire article text instead of just the article title. The data we collected required a lot of cleaning.

2. We needed to decide how to transform natural text into numbers to model the evolution of text as a dynamical process. The method of embedding words/phrases into vector space via neural networks is not applicable because of the long training time. We decided to use semantic triples, which can be extracted quickly. Furthermore, semantic triples from a series of articles collected over time can be viewed as a sequence of RDF graphs. The downside of semantic triples is that the extraction is also very noisy and may not convey the exact information from the original text.

3. We had to decide on how to model the dynamics of RDF graphs. Dynamical model of graphs is an open research area. Furthermore, RDF graphs contain semantic information on both nodes and edges. To simplify the problem, we only consider the graph structure. Instead of modeling the change of the entire graph structure, which would require a complex and high dimensional model, we decided to use simplifier graph features. The stochastic changes of these graph features were then modeled using a discrete-time, multivariate Hawkes process to account for nonhomogeneous dynamics.

Through experiments, we showed that characterizing the dynamics of RDF graphs gives more information than looking at article counts alone and that there is a difference (both in publication dynamics and semantics) between articles published from major news media and from celebrity gossip news outlets in the Jussie Smollett story.

We also showed that the discrete-time, multivariate Hawkes process can be used to model the dynamics of how RDF graph structure changes over time. However, this is preliminary work.

Development environment consisted of Python 3.7.4 from Anaconda 1.9.12 which contained all Python-related dependencies (specific libraries included: EventRegistry 8.7, Newspaper3k 0.2.8, NeuralCoref 4.1.0, spaCy 2.3.2, NLTK 3.4.5, Gensim 3.8.1, and torch 1.5.1).

6.0 RECOMMENDATIONS

There are many things we would like to do for future work:

- We have spent most of our time cleaning and analyzing the Jussie Smollett dataset. We would like to extend our approach to the other dataset we've collected and collect more data of different types of events.

- We would like to collaborate more with NLP experts to better extract/clean triples and build RDF graphs that better reflect the natural text.

- We would like to investigate more sophisticated parameters learning methods for the Hawkes process and the predictive capability of Hawkes process on the dynamics of news events. While it is generally impossible to predict how news events will unfold, it seems feasible that the 'normal' reduction of news articles over time may be predictable.

- Lastly, while we presented our method in term of news evolution over time, evolutionary need not be strictly in the temporal sense. For example, we can consider a single but very long text and build a sequence of RDF graphs (possibility corresponding to each paragraph or chapter) to characterize how the text evolved over the writing process.

7.0 REFERENCES

[1] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.
[2] R. Rehurek and P. Sojka, "Gensim--python framework for vector space modelling," NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, vol. 3, no. 2, 2011.
[3] A.-L. Barabási, "Network science," Barabási, A. L. (2013). Network science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 371, no. 1987, 2013.
[4] D. Easley and J. Kleinberg, Networks, crowds, and markets, Cambridge university press, 2010.
[5] P. J. Laub, T. Taimre and P. K. Pollett, "Hawkes processes," arXiv preprint arXiv:1507.02822.
[6] M.-A. Rizoiu, L. Young, M. Swapnil and X. Lexing, "A tutorial on hawkes processes for events in social media," arXiv preprint arXiv:1708.06401 (2017).
[7] J. Etesami, N. Kiyavash, K. Zhang and K. Singhal, "Learning network of multivariate Hawkes processes: a time series approach," in Thirty-Second Conference on Uncertainty in Artificial Intelligence, 2016.
[8] R. Browning, D. Sulem, K. Mengersen, V. Rivoirard and J. Rousseau, "Simple discrete-time self-exciting models can describe complex dynamic processes: a case study of COVID-19," Plos one, vol. 16, no. 4, 2021.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

HTML	Hypertext Markup Language
KB	Knowledge Base
MLE	Maximum Likelihood Estimator
NLP	Natural Language Processing
RDF	Resource Description Framework
REL	Relation Extraction and Linking
URL	Uniform Resource Locator