Measuring Responsiveness in the Online Public Sphere for the 2016 U.S. Election

Pau Perng-Hwa Kung  
Media Lab  
Massachusetts Institute of Technology  
Cambridge, MA 01239  
pernhwa@media.mit.edu

Abstract

The election narrative is formed under the competitions of ideas among critical players involving politicians, news media, public influentials, and the general public. Untangling the complex process of narrative formation, however, is no easy task due to implicit influences among the key players. In this paper, we propose the problem of measuring "responsiveness" that quantifies a player's influence on another given a specific election topic over time. In particular, we make use of multivariate Hawkes Process to infer the influence network between pairs of election players. We demonstrate an early version system of analytic pipeline of online public sphere discussions from data ingestion, influence inference, to visualization. The paper concludes by showcasing some preliminary results based on Twitter and news media election-related contents from July to October 2015 and discussing plans for future research.

1 Introduction

The 2016 U.S. Election is quite different from the previous elections: we have a more mature online public sphere than ever before. Recalling the German philosopher Habermas, the public sphere is an area where individuals gather to freely discuss on public issues[5]. With the prevalence of online social media, people’s opinions on public issues have become more transparent[3], and we believe it is now possible to track the formation of narrative on a given election topic.

However, the dynamics that govern the election narratives are complex: in an election, there are politicians, news media, influential public, and the general public making up the campaign voices. Each voice wants to get the podium to express its own views and interests. In other words, the election process is essentially a competition of ideas. At any given time, each voice may have varying influence on another regarding a certain election topic. In order to better understand the political narrative, it is important to be able to determine who and what that lead to a particular voice’s discussion on a certain topic. During a specific time window, is it possible to identify key opinion influencers and triggers that affect a voice to respond differently on an election topic? Developing a computational understanding can provide the general public a clearer picture on which politician or news organization has aligned topic interest. It is informative to know are the politicians responding on issues out of genuine care or are they simply acting in attempt to stand out among fellow competitors.

In this paper, we develop a computational measure of narrative influence in the online public sphere, called "responsiveness". Essentially, we model each voice’s reactions on a particular election topic as a point process, where a voice can be an individual, organization, or a group of individuals with aligned interest. Jointly considering voices as multiple point processes, we model the cross-voice influence using an excitatory point process called multivariate Hawkes Process. We describe a solution
framework entailing our data collection process, algorithm inference, and outcome visualization. A quick snapshot of the initial version of the system is shown, focusing on election-related tweets and news articles collected over the period of July thru October in 2015. We illustrate some initial application areas using the inferred influence between voices over time including interest alignment analysis, situational analysis, and topic shift analysis. We conclude by discussing potential improvements of the current algorithm and application cases.

2 Problem and Solution Framework

2.1 Preliminaries and Problem Formulation

Suppose given a specific content filtering criteria \( C \), where the filtering criteria may be discussions regarding certain election topics (e.g. abortion) or candidates. And let the number of election voices we wish to analyze be \( K \). We observe \( K \) series of events, where each series \( i \) represents event process belonging to voice \( i \). Events can be represented as a finite sequence of marked tuples \( \{s_n, c_n\}_{n=1}^N \), where \( c_n \in \{1, \ldots, K\} \) denotes the process in which the \( n \)-th event occurred and \( s_n \in [0, T] \) denotes the timestamp of the event.

We define the measure Responsiveness \( R : N \times N \times H \rightarrow \mathbb{R} \) to be the following: responsiveness \( R(s_i, t_i, h(C)) \) is a function that measures influence from process \( s_i \) to process \( t_i \) given data filtering criteria \( C \), where \( s_i, t_i \in S, K = \{1, \ldots, K\} \) denote the indices of the influencing process and the influenced process, respectively. \( h \subseteq H = [1, T] \) is the time interval in which we want to measure the responsiveness. Therefore, the problem we face in this paper is that given any sequence of comments by a set of voices regarding any election topic, can we infer Responsiveness for pairs of voices over a certain time period?

In fact, this measure can be further augmented to measure nuanced signals associated with the process-to-process influence. For example, we can add an augmenting variable \( \Gamma = \{1, \ldots, Z\} \) denoting a set of election topics (e.g. \{abortion\}, \{immigration\}, \{gun control\}) so the Responsiveness can now be \( R(s_i, t_i, h, \gamma(C), \gamma) \), \( \gamma \in \Gamma \).

An example would be to let \( C \) = discussion on immigration; \( \gamma \) = [Mexican border]; \( s_i \) = Donald Trump; \( t_i \) = Jeb Bush, so measuring the Responsiveness between Trump and Bush over time period \( h \) would also mean how much Donald Trump’s influence on Jeb Bush to react on immigration is due to cues based on Mexican border.

2.2 Multivariate Hawkes Process

To model responsiveness, we borrow ideas from a general Point Process model called Hawkes Process. In particular, a multivariate Hawkes Process consists of \( K \) point processes where the sequence of marked events \( \{s_n, c_n\}_{n=1}^N \) have direct correspondence on the process in which the \( n \)-th event occurred. For each process \( k \), it is a conditionally Poisson process where there is an associated rate \( \lambda_k(t)\{s_n : s_n < t\} \) based on the event history up to time \( t \).

We can further breakdown the rate function \( \lambda_k(t) \) into two components: one based on the process \( k \)'s own history, also called the background rate function, in which we denote as \( b_k(t)\{s_n : s_n < t \land c_n = k\} \). In this paper, we view \( b \) as a constant throughout time, which is essentially the same as the background rate for Poisson process.

The other is non-negative influence based on the event history of other point processes \( k' \) within a time window of influence \( \Delta t \), in which we denote as \( \rho_{k', k}(t)\{s_n : t - \Delta t < s_n < t \land (c_n = k \lor c_n = k')\} \).

Figure 1 illustrates a sample Hawkes process. Whenever an event occurs at process \( k \), we place a window of impulses on other processes \( k' \) via \( \rho_{k', k} \) if \( \rho_{k', k} \) is nonzero. Superimposing the background rate \( \lambda_k \) (denoted by flat horizontal line), we have a combined influence leading to event occurrence in any process \( k \) over time.

In detail, we can further decompose \( \rho_{k', k}(t) \) into \( W_{k', k} g_{k', k}(t|\Delta t) \), where \( W \in \mathbb{R}_+^{K \times K} \) is a non-negative weight matrix denoting influence from process \( k' \) to \( k \). \( g \) is a convex combination of basis
functions, each curtailed to be of length $\Delta t$ with nonzero values, so $g$ is a convolved signal in time with the basis functions.

The joint likelihood of the model is the product

$$p\{s_n, c_n\}_{n=1}^N | \lambda_k(t) = \prod_{k=1}^K p\{s_n : c_n = k\} | b_k(t) \prod_{n=1}^N \prod_{k=1}^K p\{s_n' : c_n' = k\} | \rho_{c_n,k}(t - s_n).$$

For parameter learning, we use stochastic gradient descent on the log-posterior function, where we have placed a Gamma prior on the weight matrix. To use the outcome of the algorithm, we compute instantaneous responsiveness $R(s_i, t_i, t|C)$, $t \in [0, T]$ to be

$$R(s_i, t_i, t|C) = W_{s_i,t_i} \int_{\delta t=0}^{\Delta} |\{s_n : c_n = t_i \land s_n = t - \delta t\}| g(t - \delta t) dt$$

To represent responsiveness over time interval $h$, we simply integrate over instantaneous responsiveness rate and normalize by receiver $t_i$.

### 2.3 Solution System Framework

As we build towards an end-to-end application system using inferred responsiveness, we want a system from data ingest to output visualization. Figure 2 shows the current flowchart and application
snapshot of system design. The first part of the system consists of realtime data ingest from Twitter and 13 mainstream news sources. An NLP post-processing component is used to filter tweets and news articles that are election relevant. The next part contains voice and topic identification to filter contents and involving voices by user’s interest.

The main component of the application is the inference of responsiveness. Based on the instantaneous responsiveness \( R(s_i, t_i, t_i|C) \), we can plot influence of other voices on a voice of interest over time. We can also examine the underlying weight matrix \( W \) as an influence network. To check the exact cues that lead to influence, we enable the functionality to examine actual contents associated with the events.

3 Related Work

There are two lines of related work for measuring responsiveness in the online public sphere: analysis of influence and dialogue from political theory; computational modeling of multiple time series influence.

3.1 Political Influence Analysis

The first line of work is about political discourse analysis in the political theory. Political discourse analysis looks at how uses of rhetoric, lexicon, expression structure, interactions among the actors can shape political narrative[2]. It serves as a qualitative framework for analysis, and less of a quantitative methodology for measuring influence.

3.2 Multiple Process Dependence Modeling

Modeling dependence between multiple time series relates to the concept of Granger Causality[4]. Granger Causality is a regression-based hypothesis test of causality between pairs of time series. However, in general, we would like to understand how each pair of voice influence jointly relates to one another as well as the nuances leading to different influence effects, which require capabilities beyond simple hypothesis testing.

We can also view each time series as a discrete sequence of events. Recently, variants of Hawkes process have been proposed in the machine learning community to model latent dependencies among a set of points processes[1][6][9]. The Bayesian nature of the models and their flexibility for generative hypothesis make this family a good option for first-pass design.

4 Initial Applications

4.1 Data Collection

Based on the initially implemented version of application, we are taking the ingested data from the data pipeline for preliminary analysis. For news articles, we have a realtime article ingest of 13 mainstream news organizations(AP, Buzzfeed, CNN, Fox News, LA Times, McClatchy, New York Times, NPR, Politico, ProPublica, Reuters, Wall Street Journal, Washington Post). For tweets, we include all tweets posted by news organizations, presidential candidates, and any tweet with text matching a predefined query set. We use a particular set of topics: [abortion], [Mexican border], [immigration], [gun control], [Syria], and [terrorism].

4.2 Influence Analysis

For preliminary analysis, we test on three potential applications: interest alignment analysis, situational analysis, and topic shift analysis.

4.2.1 Candidate-Candidate Influence and Alignment of Interest

If we let the voices \( s_i \), \( t_i \) each represented by a presidential candidate, we can use the candidate-to-candidate responsiveness to understand topic interest alignment between the candidates. Figure
Figure 3: Inferred responsiveness showing the influences of other presidential candidates on Jeb Bush, making Jeb Bush tweeting about Mexico (top), immigration (bottom) at given time.

3 shows inferred responsiveness over time for presidential candidates influencing Jeb Bush to tweet about Mexican border and immigration. When we navigate through the time (x-axis), we see spikes on the y-axis. These spikes represent strong influences from certain candidates. If we are interested in how Donald Trump influences Jeb Bush’s narrative, we see that this is only the case for immigration, whereas Trump has very little influence on Jeb Bush over Mexico debate. This reveals an interesting phenomenon of the immigration debate among the Republican candidates: because Donald Trump was making inflammatory comments about Mexican immigrants, other candidates were consciously avoiding Trump’s narrative on Mexico and they instead focused more on the general immigration debate.

4.2.2 Media’s Preference on Candidates

It is also interesting to look at the responsiveness of news media influenced by presidential candidates’ comments on Twitter. Figure 4 shows the candidate influence on Fox News to post articles on Mexican border and immigration respectively. We can observe the contrast how Donald Trump strongly influenced Fox to write about both Mexican border and immigration. Also, we can see how Fox news lowered its interest on the both topics after September began.

4.2.3 Situational Analysis

Lastly, the responsiveness can be used for situational analysis to compare the differences of aligned interest before and after a certain news event. Take Figure 5 (left) as an example, the chord diagram shows the candidate influence on Hillary Clinton tweeting about abortion before and after the first Republican debate. Before the debate, we see a big proportion of influence from Rand Paul and Scott Walker, who were both very active on the topic. However, after the debate, there was a big drop of influence from the two candidates. If we referred to the Republican poll[7], we knew the two candidates’s support dropped substantially post debate. It seemed natural for Hillary Clinton to shift focus towards more competitive Republican candidates on the issue.

Another example of situational analysis is the analysis on shift in topical influence over time. Figure 5 (right) displays how the topical influences that led to the differences in news media’s coverage on immigration varied before and after the beginning of September. Interestingly, the media coverage on Mexican border dropped significantly after the Republican candidates visited the border in late August. Therefore, there was less influence from the topic Mexican border on immigration starting from September.
Figure 4: Inferred responsiveness showing the influences of presidential candidates on Fox News writing about Mexico (top), immigration (bottom) at given time.

Figure 5: Illustration of (left) candidate influence on Hillary Clinton tweeting about abortion before and after the first Republican debate; (right) topical influence on news media writing about immigration before the after the beginning of September.

5 Discussions and Future Work

In this paper, we have proposed to use multivariate Hawkes Process to model influence dependence between different voices, or point processes, over time. We have implemented an initial version of the end-to-end system that takes the ingested data for influence analysis in the online public sphere. We demonstrate a few use cases where our responsiveness can deliver interesting analysis, where the outcomes may be used as part of journalistic reports discussing the opinion influence dynamics in the public social media ecosystem.

Although the preliminary analysis suggests some interesting potential application using the responsiveness measure, this work is still very preliminary in general. There are several important hanging challenges that need to be resolved in order to make the output of the responsiveness measure more useful:
• The current model is not very good at handling different types of voices that might have mixed scales of magnitude. For example, the sequence on general public tweeting about a topic will have much higher magnitude than presidential candidates. How can we unify the effect of popularity in each voice, beyond simple normalization, so the inferred responsiveness is not easily dominated by the difference in scale (by placing assumption on virality of a topic, as described in [8])?

• The current model does not generate a sparse set of positive weights. When we include sparsity in the model, can we still effectively retain the explainability of inferred influence, or will the inferred influence dominated by sparse generative assumption?

• Ultimately we would like a model that is able to infer nuanced factors leading to certain reactions. How might using "Big, Beautiful Wall" on the Mexican border topic have a different influence on people reacting than using "illegal immigrant"? We need to expand the generative model to incorporate such data attributes.

Acknowledgements

The authors would like to thank Sophie Chou, Neo Mohsenvand, Prashanth Vijayaraghavan, and Soroush Vosoughi for suggestions on research, data collection, and providing article data. The authors also thank Ryan Adams, Allen Gorin, and Scott Linderman for providing resources and suggestions on algorithm design.

References


