Understanding Information flow using different perspectives

HUSSAM HABIB, University of Iowa

1 INTRODUCTION

Information flow is the process of information propagating in a social network through its users. An example of understanding information flow would be to track the propagation of a meme on an online social media. We can study the dynamics of information flow on social media by studying the rate at which information gets propagated and the structure of the cascade it forms. In this report we look at how different variables impact information flow. We can observe multiple variables to understand information flow. These variables can be divided into 5 perspectives: 1) **content features**, 2) **user features**, 3) **structural features**, 4) **temporal features** and 5) **platform features**.

Current social media platforms design has enabled them to use algorithms to provide users with information that is relevant to them. By doing this the information reaches user without their direct involvement in what they want to actively seek out [15]. This change in how information reaches users has drastically changed how the population gets their news, entertainment, and advertisements and how the users are affected by it. Current social media platforms differ with each other in the content they provide, how the users are linked to one an other and how information reaches the users. Generally social media can be viewed as a network of users as nodes and edges as paths which information can take to reach these users. As discussed before information flow is the study of understanding the dynamics of how information spreads in this network.

Anecdotal evidence and multiple case studies show online social media as integral tools in creating and spreading information including conspiracies, misinformation, disinformation and propaganda [16]. One of these cases is QAnon [7] which is a conspiracy theory started on 4chan that President Donald Trump is facing down an evil cabal of democratic pedophiles. Since its inception on 4chan it has gain enough popularity to have made appearances on mainstream news domains and channels. There are many similar cases of conspiracy theories, misinformation and ideologies that stared from fringe communities and end up in mainstream news such as the Boston bomber incident [1], the pizza gate conspiracy [2] and the GamerGate movement [3]. These cases show Information has the ability to manipulate user behavior and information cascades can scale this affect on a large number of users [13], [12] and [14]. The importance of understanding the dynamics of when and how information cascades succeed in relation to the initial configurations can be divided into two use cases. 1) Having understanding and control over parameters that can make information cascades successful can help marketers, political activists, propagandists, influencers and government to spread their message to a large number of users. On the flip side, 2) Understanding the mechanism on how information cascades succeed can aid platform moderators to intervene such malicious cascades and mitigate their spread in communities.

Author's address: Hussam Habib, University of Iowa.

#	Study			Methodology			Perspectives				
	Name	Year	Venue	Analysis	Key Technique	Platform	Content	User	Structure	Temporal	Platform
1	**Web centipede**	2017	IMC	Empirical	Hawkes Processs	Twitter, Reddit and 4chan	✓		✓	✓	✓
2	**Generating Realistic Interest-Driven Information Cascades**	2020	ICWSM	Generated	Generating cascades	Simulation	✓	✓	✓		
3	**Can cascades be predicted**	2014	WWW	Predictive	Logistic Regression	Facebook	✓	✓	✓	✓	

Table 1. Selected paper techniques and perspectives used.

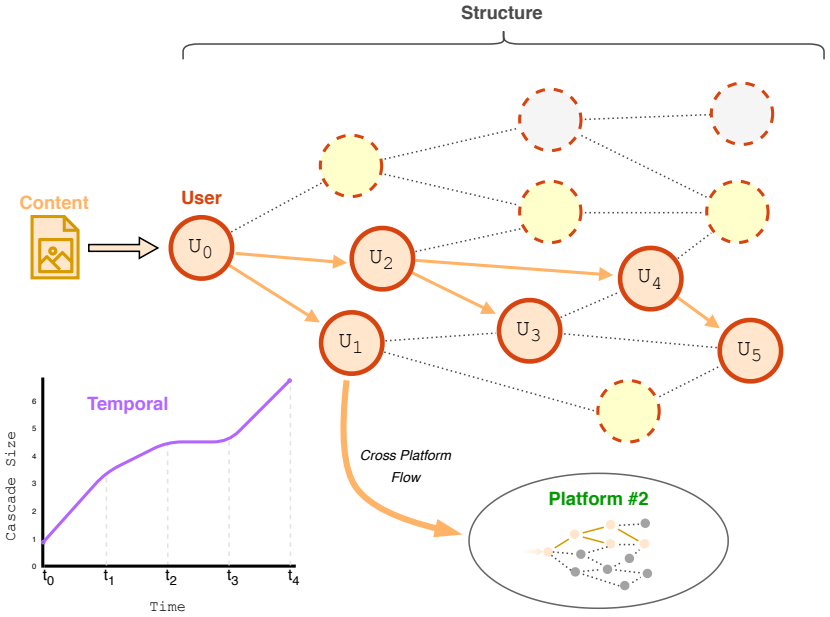
1.1 Objective

We look at research done in the field of measuring and understanding information flow on online platforms. These works aim to measure and characterize information flow, understand the variables that can affect it and measure existing cases of information flows. Each paper uses a different perspective in understanding how different variables affect the propagation of information.

(1) **The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources**. The authors of this paper [18] look at existing social media platforms and measure the propagation of mainstream and alternative news URLs. They perform an empirical analysis and use measurement techniques to characterize each platform based on the propagation of these URLs, study the influence of platforms on each other and how the different combinations of information (types of URLs) and platform can affect the size of the cascade. Their perspective of understanding information flow includes the type of content, features of the user base, characteristics of the platform, and the influence of each of the platform.

(2) **Generating Realistic Interest-Driven Information Cascades**. The authors of this paper [10] model a social network and generate information cascades to understand how different variables in the their social network model affect the size and structure of the cascade. The variables they focus on are user interests, user influence, content topics and the structure of the network. Generating information cascades by modifying these variables the authors reveal how each of the variable affect the final cascade structure.

(3) **Can Cascades be Predicted?**. The authors design a train a machine learning model to predict whether a cascade will grow big or not [9]. By identifying cascades on Facebook and extracting features from these cascades the authors successfully train a logistic regression model proving the predictability of cascades. The authors focus on user features, content features, cascade structure, and temporal features to train their model. Furthermore, the authors also explore how each of the feature affect the predictability of the model consequently exploring how each of the feature affect the size and structure of the cascade.



Perspectives	Variables
Content	Virality, Language features, original source, content type, topics, etc.
User	Number of followers, is a bot, influence, interests, age, activity etc.
Temporal	Rate of propagation, time between reshares, time between views, etc.
Structure	Average connection count, homophily, cascade depth, out degree, etc.
Platform	Platform influence, size of user base, platform policies, type of social media, etc.

Fig. 1. How information propagates in a social network.

1.2 Background

In this section we describe information flow and the different perspectives we are studying. In the Fig. 1 we show how information propagates in a network of users connected with each other. The orange arrows and users show the flow of content. The yellow users represent users who saw the content but did not share it further and users who never received the content are colored gray. The perspectives we look into to understand the dynamics of cascade evolution are also labeled with a few examples provided for each of them.

2 THE WEB CENTIPEDE: UNDERSTANDING HOW WEB COMMUNITIES INFLUENCE EACH OTHER THROUGH THE LENS OF MAINSTREAM AND ALTERNATIVE NEWS SOURCES

This paper studies propagation of mainstream and alternative news URLs within and between Twitter, Reddit, and 4chan.

2.1 Objective

This paper studies and measures the information flow in terms of alternative and mainstream news URLs on three different online platforms Twitter, Reddit, and 4chan. These platforms are selected because of their fundamental differences in structure and the impact these platforms have on forming and manipulating people's opinions. The authors perform a three level analysis to understand how these URLs flow within and between the online platform and create an online ecosystem.

(1) **Characteristics of each of the platform.** By measuring the occurrences of mainstream and alternative news URLs separately they create a general characterization for each of the platform in terms of the ratio between the two types of the URLs and the ratio of users who share these URLs.
(2) **Temporal dynamic each platform shows during a cascade.** In their next analysis they start to track each URL within a platform over time helping them understand the rate of information propagation. This analysis helps them understand how the speed of information flow is different for each of the platform and how the different types of URLs change this rate of information propagation.
(3) **Influence each platform has on each other.** Finally to understand the entire process of information propagation online, they set out to study the online ecosystem these three platforms create together.

By measuring the, occurrences, temporal dynamics and the rate of information propagation between the three platforms they create a more complete picture of the path these URLs take online. In the final analysis they study the influence each platform has on each other and how that influence changes with respect to the type of URL being shared. In conclusion the authors measure the flow of mainstream and alternative news URLs within and between three social media platforms and uncover how the differences between the platforms and the different type of information can affect the flow of information. In more detail, they look at how the source and type of a contagion, the fundamentals of a social media platform, the size of the user base and the influence platforms have on each other can dictate the dynamics of the information flow. Therefore, this project helps us add new perspectives and variables to the our collection of factors that impact information flow.

2.2 Problem Statement

The authors perform empirical analysis to track mainstream and alternative URLs on three different platforms and understand how the different information propagates differently on each platform. This opens an opportunity to observe how these differences impact the information flow.

2.2.1 Measuring impact on information flow. The information flow path and the social network as a whole can be represented as a network structure. In this network each node represents a user and each edge represents a directional flow of information between two users. Using this network representation a case of information flow can be quantified. The authors use the depth, width and total size of the information flow network to measure impact on information flow. The empirical analysis the authors perform to understand the impact of variables on information flow can be divided into two categories: 1) Variables that are inherent to information and 2) variables that are related to the network structure and the fundamentals of the platform.

2.2.2 Platforms. The authors select three different platforms to track information on. In this section we describe each of the platform and highlight on their differences.

(1) **Reddit** can be described as a collection of forums. More commonly Reddit is known as a social news website and forum where content is socially curated and promoted by site members through voting. Users can subscribe to different forums known as subreddits to get popular content from those subreddits show up on their homepage and also go and visit those subreddits. Users

can make submissions also post comments to discuss the content in the post. Many subreddits discourage reposting posts made by users therefore decreasing the instances of resharing the same information multiple times.

(2) **Twitter** follows are more word of mouth social media structure. Each user has the choice to follow other users and have the content they have posted show up on their newsfeed. The users can then like, comment or retweet the tweet which would propagate the tweet to all their followers. This social media follows the information flow as a tree structure more closely.

(3) **4chan** is similar to Reddit where it is collection of forums where users can make posts and comment about that post in relevant forums. In this study however the authors look at the /pol/ the purpose of which is to discuss news, world events and political issues. The main difference in 4chan is that it is anonymous and ephemeral. Comments and posts can not be traced back to those who made it and will be deleted in a few days. Because of the anonymity, ephemerality and lack of rules content on 4chan can become very toxic and hateful.

2.3 Understanding information flow from different perspectives

The authors use three perspectives to understand information flow **Content**, **Temporal** and **Platform**.

2.3.1 Content. To understand how content can impact information flow they track the flow of two different categories of URLs having different sources. The authors track the flow of URLs from mainstream news domains and URLs from alternative news domains.

Information variables. The authors categorize these URLs as either from a mainstream news source or alternative news source. Alternative news sources include state sponsored news sites as well as commonly known fake news sites that propagate propaganda. Furthermore the ubiquity is of these domains (mainstream URLs being more recognizable and known) is also a key difference between the two type of the news URLs. Finally, although this might just be an assumption it is more likely for mainstream news URLs to uphold journalistic integrity while alternative news sources might bend journalism rules and morality just for the sake of spreading their content more efficiently and therefore trying to make content that is more biased, controversial and viral [8]. It also might be considered, that although this is not the inherent quality of the information that is being spread, the initial spread/propagation of alternative news sources is different compared to mainstream URLs, where it was seen in this project that alternative news sources cascades were more likely to begin with bots or users who exclusively spread alternative news URLs (which can be seen that their main motive is to spread alternative news rather than just a piece of information for the sake of sharing it which might be the case for many mainstream news). When observing different types of information this study does not get differentiate between information other than only looking at their sources as either alternative news domains or mainstream news domains.

2.3.2 Temporal. In their analysis the authors also look at rate of information propagation and how it affects the reach of information within and across platforms. They look at how news cycles are different in each of the platform. Furthermore, hey also look at the average time between subsequent shares in different platforms. They also look at mean inter arrival time when looking at cross platform propagation.

2.3.3 Platform. The authors perform their analysis on three fundamentally different platforms. The authors highlight the differences between each of the platform and show how these difference result in different cascade sizes for different type of information. How these three social media are setup set them apart significantly. Looking at these differences will help us understand the variables which help information spread differently in these communities.

(1) **Homophily** One other variable that we see engender due to these platform difference is the sense of homophily and ease of forming echo chambers. Twitter can form homophily by users only following users with similar interests and beliefs [11], however Reddit is more prone to homophily and echo chamber effect because of the fact that users can follow communities (for example The_Donald) where the guidelines discourage non-conforming content and content that gets upvoted to the top in a community is usually the content voted by an already biased set of users in the community [17].
(2) **Bots and organic shares** The use of bots to tweet and retweet content on Twitter to inorganically increase its visibility plays a role in propagating information as well, especially news from alternative news sources [6]. There are also many state sponsored users on twitter that spread propaganda.
(3) **User Base** Twitter is clearly is the most popular platform with the highest number of users, followed by Reddit and finally 4chan. Furthermore, there is a difference of demographics of users active on Reddit, 4chan and Twitter /citereddit-demographics, [4]
(4) **User interface** The ease of sharing posts and causing cascades on Twitter in unmatched by any of the other two platforms. Twitter allows users to retweet posts using a single button.
(5) **Cross platform influence** Another analysis the authors perform is to understand how information flows between the platforms. They measure the influence each platform has on the other. By looking at this they can understand how the factor of influence between the platforms can play out.

2.4 Methodology

The authors perform a three tier analysis of each platform to explore the differences in each of the platforms and how it is affecting the current online ecosystem in terms of spreading news from mainstream and alternative sources. The authors perform a three step analysis as described below:

(1) **General Characterization.**They attempt understand how the fundamental differences in structure, network and user base of these platforms affect the spread of information. They perform a simple study to measure the occurrences of mainstream and alternative news sources, this analysis measures the instances of external or internal actors initiating the information flow.
(2) **Temporal Dynamics Within Platforms.** This analysis focuses on information flow on the same platform, they measure and analyze the time it takes for each type of URL to spread on a particular platform. This analysis helps understand the speed at which information flows internally in the platform and how the differences between the structure of platform, user base and type of content impacts the delays between the process of re-sharing.
(3) **Temporal Dynamics Cross Platform.** They explore the information that flows outside the platforms and between other platforms. They measure the time delays between instances of the same URL showing up on two different platforms and by measuring this on a large scale they

understand the influence each platform has on the other. This analysis introduces a new variable of inter social media influence that can impact the speed at which information flows between social medias.

2.4.1 Data Collection. The authors select Twitter, 6 communities from Reddit and the /pol/ board from 4chan to study as platforms. The authors select these platforms because of the popularity in disseminating mainstream and alternative news and the differences between the network structure and user base of these platforms. The authors collect 1% of the public tweets from twitter between the dates of June 30, 2016 and February 28, 2017. The authors collected data from the top 20 most popular communities on Reddit from dates June 30, 2016 and February 28, 2017 to perform the first analysis and used selected 6 communities r/The_Donald, r/politics, r/conspiracy, r/news, r/worldnews, and r/AskReddit.

The news URLs that the authors track are originate from two different sources, mainstream and alternative. URLs from a total of 99 news domains are tracked from which 45 are mainstream and 54 are alternative. For collecting mainstream news domains they select the top 45 sites ranked by Alexa.

The authors use data from Twitter, 4chan (/pol/)and 6 communities from Reddit to explore the flow of URLs from 99 (45) mainstream and (54) alternative news sites. They use posts made on these three platforms collected from July 2016 to March 2017 and keep only posts that contain URLs from one of the 99 news sites.

2.4.2 General Characterization. The goal of this analysis is to characterize each of the platform based on the distribution of different types of URLs and the sources of those URLs. This process is simply done by measuring the occurrences of these URLs and their news site domains. They also perform further analysis by measuring the fraction of alternative and mainstream URLs per user.

2.4.3 Temporal Dynamics Within Platforms. In this section the authors start looking at the flow of information within the platform. They measure the time delay between the initial act of sharing a URL and the subsequent sharing of the same URL. This analysis helps them understand the speed and depth at which information disseminates on each platform. In understanding the reposting behavior of each of the platform they again simply measure the occurrence of each unique URL that occurs more than once in the platform along with the time stamp of when it was posted. This provides them with all the raw information from which they can draw conclusions on how quick the flow of information is.

2.4.4 Temporal Dynamics Cross Platform. In this section the authors look at information flowing out of a platform and into another. This analysis enables the authors to understand the online ecosystem these three platforms have created and measure the influence each platform has on the other. To measure the influence of a platform they model a Hawkes Process with a process for each of the platform. . An event is considered the posting of a URL. All the communities are considered a process, 1 process for Twitter, 1 for 4chan /pol/ board and 6 for each of the 6 subreddits from Reddit. There are two ways this event can occur, due to a background process or due to one of the other three process. A background process is modeled after the probability of the event (posting a URL) occurring without any influence from any of the other 8 processes, this include the user posting the URL for the first time in the ecosystem directly from the source or from one of the other platforms not modeled. Between each pair of the process P_i and P_j there is a weightage that reflects the probability of the event occurring in P_j due to the event occurring in Pi. This weightage can be seen as the influence P_i has on P_j on the event type. The

Hawkes process also models the concept of self influence which is the probability of an event occurring on Pi because of an event already occurring on the Pi. This can be seen as a person retweeting a tweet on Twitter. The authors model 2 Hawkes processes, one where the event is the posting of a mainstream news URLs and one where the event is the posting of an alternative news URL.

2.5 Results

The authors use the aforementioned methodologies to explore how information flows in each platform differently for each type of content differently. This unveils an opportunity to see how the differences in the content and platforms reflect in the flow of information.

2.5.1 General Characterization. The authors first task is to create a general characterization for each of the platform for each type of URL separately. In creating this characterization the authors measure the occurrences of each URL for each platform the authors results explores the initial stages of understanding information flow in terms of the information's spread and reach. The authors observe the following results in their analysis:

- Alternative domains occur on the three platforms equally, however, the distribution of different domains of alternative news URL is different on each platform with some domains
- The six subreddits have the most contribution in disseminating controversial news.
- Alternative and mainstream news get the same number of retweets. 80% of users share all he mainstream news URLs however only 13% share all of the alternative news URLs. This can be attributed to bots driving many alternative URLs on twitter. Furthermore, it is very unlikely for a user to never share an alternative news URL.
- 4chan has the same distribution of mainstream and alternative news URLs. However some alternative news domains appear exclusively on 4chan.

2.5.2 Temporal Analysis within platforms. In the second analysis of how information cascades are different in each platform the authors include the temporal dynamics in the measurement.

- All of the platforms had statistically different distribution of mean inter arrival time. Reddit had the highest and Twitter had the lower. Twitter also had the smallest lag between the first and subsequent occurrences.
- Compared to mainstream news URLs, alternative news propagates cross platform at a much higher rate. However, for propagation within the platform mainstream had a higher rate than alternative news.
- Cross platform propagation depends on the combination of source platform and content type. Some platforms, such as Reddit (specifically The_Donald) are more successful at spreading alternative news to other platforms. Furthermore, Alternative news cross platform propagation is fastest when the destination is Twitter, this can be attributed to the large and diverse user base on Twitter [5].
- For URLs that appear on all three of the platforms (Reddit: R, Twitter: T and 4chan:4) : R->T->4, R->4->T, T->R->4 are the most common sequences in order.

2.5.3 Measuring Influence. Finally the authors measure the influence each platform has on the other for different different types of URLs. They model two Hawkes Processes one for each type of URL and

look at the weights between each of the platforms which is equivalent of the influence each platforms has.

- Twitter has the highest background process rate which means it is more likely for a Twitter user to post a URL for the first time and without any influence from any of the 8 process (platforms).
- The background rate for alternative URLs was higher than for mainstream URLs on all platforms.
- Twitter also had a highest influence on other platforms for mainstream news URLs. Out of the 6 subreddits on Reddit, The_Donald was influenced by other platforms on mainstream news URLs and had the highest influence on other platforms for alternative news URLs. This suggests The_Donald users are exposed to both mainstream and alternative news at the same rate but are more likely to share alternative news on other platforms.
- Twitter, being the biggest and most popular platform, is influenced by fringe communities (The_Donald and 4chan). 6% of mainstream and 4.5% of alternative news URLs posted on twitter come from these communities.

3 GENERATING REALISTIC INTEREST-DRIVEN INFORMATION CASCADES

In this paper the authors model a social network as a graph with nodes as users and edges as the connections between users. They use this graph to propagate information in the network by modifying user interests, user influence and content topic.

3.1 Introduction

In the first paper (Web Centipede), the authors performed an empirical study on existing social media platforms to track information flow. In this paper the authors highlight how existing platforms are hard to measure because of limited access to data and non idea conditions. They create a social network model called WoMG (Word of Mouth Generator) which is a graph of users connected with each other. The objective of this paper is to use WoMG to create realistic information cascades and study how modifying different parameters in the graph changes the dynamics of information flow in terms of rate of propagation and the structure of the cascade. Using these modifications in the graph the authors are able to create causal inferences based on each of the variables. Using these inferences we can understand which variables are important for a cascade to succeed.

3.2 Problem Statement

The main objective of this paper is to model information propagation in a synthetic social network. For this task they create a model Word of Mouth Generator (WoMG). They create this model to understand how information flow is affected in two different ways, its size and depth by modifying parameters of the social network graph. To create information cascades their model takes in the following inputs: 2) a topic model to generate items to propagate in this network and 2) A directed social graph with nodes as users and unidirectional edges representing the flow of information. Using this model the authors generate realistic cascades and measure how different initial configurations drive the cascades differently. The authors look at 2 different perspectives to understand their impact on information flow: **content**, **user** and **structure**. In generating cascades that are similar to real world cascades the authors face two challenges. They have to generate the user graph with realistic user interests and realistically

homophilic networks. They also have to generate content with realistic virality and distribution of topics.

3.3 Methodology

3.3.1 Can Realistic Cascades be Generated? Word of Mouth Generator (WoMG) is the model that takes in a user graph and a list of items. WoMG uses these inputs to output a cascade vector which informs the user on how the item propagated in the user graph. Their model can be divided into two submodels: the items and the user graph. In this section we describe what these submodels are and what are the parameters the authors can tune to change the cascade.

Items. The first task is to generate items I that are to be propagated in the social network. For their model the authors list k topics for each of the item. Each item i has two parameters that can be modified:

- A topic vector γ_i of size k representing the distribution of topics.
- Virality of the item v_i.

Graph. The graph is made up of nodes, these nodes represents the users. Each edge in the graph represents a bidirectional flow of information. The edge (u, v) represents the fact that user v is a follower of users u.

Users Each user u is assigned an interest vector t_u of length k which represents the topics that they are interested in. This interest vectors determines how likely it is for a user to reshare an item, items with topic vectors similar to user interest vector are more likely to be reshared by the user. On a higher level, tuning a user's interest vector and interest vectors of connected users determines the homophily of the sub network. Finally for each of the user, the authors also assign an influence value ρ for each of the topic in the interest vector. Users with high values for ρ can be thought of as influencers.

Using the aforementioned features we can look at more complex features that emerge as a function of these simple features. The main challenge for the authors is to model these complex parameters in a way that they can be tuned by changing minimum number of parameters.

User Influence. Each of the user has an influence value ρ that represents their status in the social network, the higher this value the mode influence a user has. However, each user v exerts different influence on user u based on the topic z. $P_{v,u,z} = t_{u,z} + \rho_v \cdot tv, z$

Social Pressure. For a user u to activate on an item i at time t, the social pressure $W_i^t(u)$ exerted on them should be greater than a threshold. This threshold is the ratio of the global resistance r towards virality and the virality v_i of the item i. $W_i^t(u) >= r/v_i$. The social pressure itself is a function of user interest vector t_u, topic distribution of the item γ_i, number of users u is following who have already shared i and the pressure p each has on user u for the topic z.

$$W_i^t(u) = \sum_{z=1}^{k} \left(\gamma_i^z \sum_{v \in F_i(u,t)} p_{(v,u,z)} \right)$$

Homophily. One of the key challenges in this paper is to create realistically homophilic social networks. Homophily is an attribute that represents the diversity or lack there of in a group. Groups high in homophily have members that have similar beliefs, interests and opinions. Echo chambers are a

great example of highly hompphilic groups. In social networks we see the phenomenon of groups with high homophily quite commonly, this is usually because of the algortihmic personalization and recommendations. In this study one of the key challenges the authors face is to create social networks with realistic homophily. In terms of WoMG the authors define homophily as the ratio of average similarity among connected nodes and the average similarity among disconnected nodes. The similarity $\delta(u, v)$ between users u and v is the cosine similarity of their interest vectors. The homophily of the entire social network can be represented by the following equation

$$h_\delta(E, \bar{E}) = \frac{|\bar{E}| \cdot \sum_{u,v \in E} \delta(u, v)}{|E| \cdot \sum_{u,v \in \bar{E}} \delta(u, v)}$$

Homophily plays a role in how information propagates in a social network. To understand this phenomenon, the authors try to make homophily tunable to a wide range of values. The first step towards this is to try to come up with methods that can maximize homphily by changing the minimum number of simple features. Homphily is directly dependant on how the users are connected and what their interest vectors are. The task is to come up with parameters when generating interest vectors that can change the homophily of the social network. The authors describe and implement 3 methods to maximize and tune homophily and share the results for each of those.

Label Propagation This is the most straight forward method by which homophily can be tuned. Essentially this method assigns similar interest vectors to nodes that are connected.

(1) Select a distribution D to assign interest vectors from. And assign initial interest vectors to each of the node.
(2) Select a set of users from the graph to become influencers M.This can be done using a greedy approach where m users are selected that are the farthest from each other and are assigned as influencers. They key takeaway is that the influencers are not directly connected to each other.
(3) Update user interest vectors by propagating interest vectors of each node u to its neighbors. For each topic z the interest vector for user v is update using the following equation:

$$t_{v,z} = \frac{t_{v,z} + \alpha t_{u,z}}{\sum_k^K t_{v,k} + \alpha t_{u,k}}$$

The value for α is set in such a way that influencers never change their own interests, influencers can strongly influence interests of non-influencers and non-influencers influence non-influencers only slightly.

For this method, the parameters are the number of iterations when updating the interests and the fraction of influencers. They show higher number of fraction of influencers result in higher homophilic network up to 25% after which it starts to show diminishing returns due to fixed interests of the influencers. Unsurprisingly, homophily also increases with high number of iterations.

Matrix Factorization The goal is to generate interest vectors in a way that maximizes homophily. This problem can be rephrased into maximizing combinations of connected nodes with high similarity $\delta(u, v) \in E$ and minimizing combinations of high similarity of unconnected nodes $\delta(u, v) \in \bar{E}$. The results for this method show the most wide range of achievable homophily by tuning two parameters. Both the parameters also show a linear and positive relationship with the networks homophily.

Evaluating cascades. In this section the authors design an experiment to evaluate the fidelity of their model. Using real data sampled from Digg in 2009, the authors try to replicate the cascades found in the real data using their model. Given a different set of cascades from the real world data set with different initial configurations and final states the author can compare the coverage of cascade size and depth of the real world cascades with the coverage of cascade size and depth of synthetic cascades generated by their model by tuning the parameters.

3.3.2 How do different configurations affect cascades. In this section the authors design experiments to understand how each of the different parameters forming different initial configurations can change the eventual size and structure of a cascade. The authors select 4 macro parameters which we discuss in detail later in this section.

Measuring cascades. To understand how tuning the initial microscopic parameters affect the cascade in its later stages on a macroscopic level the authors define properties of the cascade that they can measure. The authors focus on the size and depth of the cascade. **Average cascade size** is the count of activated nodes across all of the propagated items. **Average cascade depth** is the number steps between the last and first activated node of the graph. Tuning each of the configurations and parameters the authors show how these two parameters change.

Parameter Analysis. The authors set up the WoMG with all the different parameters that can tuned and create different configurations for an initial cascade. The authors list 4 configurations by changing 4 of the following parameters

(1) **Type of initial activation** How the content is introduced to the user graph can impact how the information propagates. In real world, the content is either 1) started by a single node in a user graph as original content or someone introducing the content to the user graph for the first time or 2) the content is propagated from an external source to multiple users who then share the content starting the cascade. In this model, the type of initial activation is a binary parameter. The initial activation can be endogenous where the item is spread by a single user who has the most similar interest vector as the topic distribution of the topic. The other type of initial activation is where the model has a dummy external node that is connected to all of the users in the graph. The content is propagated by the external nodes under natural circumstances as described before.
(2) **Type of propagation** The type of propagation can either be by interest only where all of the users has no influence on each other. Alternatively, the item can be propagated by influence where both whether a user will activate on an item depends on influence and interests both.
(3) **Virality resistance** For a user to activate on an item the social pressure for that item on that user $W_i^t(u)$ should be greater than r/v_i. Here v_i is the virality of the item and r is the global constant of resistance to virality. This parameter simply determines how easy should it be for a user to propagate an item.
(4) **Homophily** This is the ratio of average similarity of connected nodes and average similarity of unconnected nodes. This shows how diverse are the sub networks in the graph. High homophily (value close to 1) suggests that users with similar interests are connected to each other and not connected to others. Homophily is synonymous to the number of echo chambers in the network.

3.4 Results

3.4.1 Can realistic cascades be generated. The main goal of WoMG is to provide the user with a minimal list of simple parameters that represent real world parameters and that the user can tune and consequently change the eventual cascade size and structure. In accomplishing this goal the authors select parameters that have the widest range of resulting cascade size and control over precision of change. In generating realistic cascades the biggest challenge was to generate realistic interest vectors that result in realistic homphilic networks. The authors select the **Matrix Factorization** method to tune homophily by assigning interest vector to users using this method. This method provides the highest range and control over resulting cascade size.

Evaluating cascades. Finally, to evaluate whether the model is able to generate cascades similar to real world cascades the authors sample cascades from Digg 2009 publicly available data and try to simulate similar cascades using their model. They show successfully that their model has the same range as seen in the real world data and can map many of the cascades. As shown in Fig. 3, with homophily at 0.5 their model is the closest to the average size of the realistic cascades.

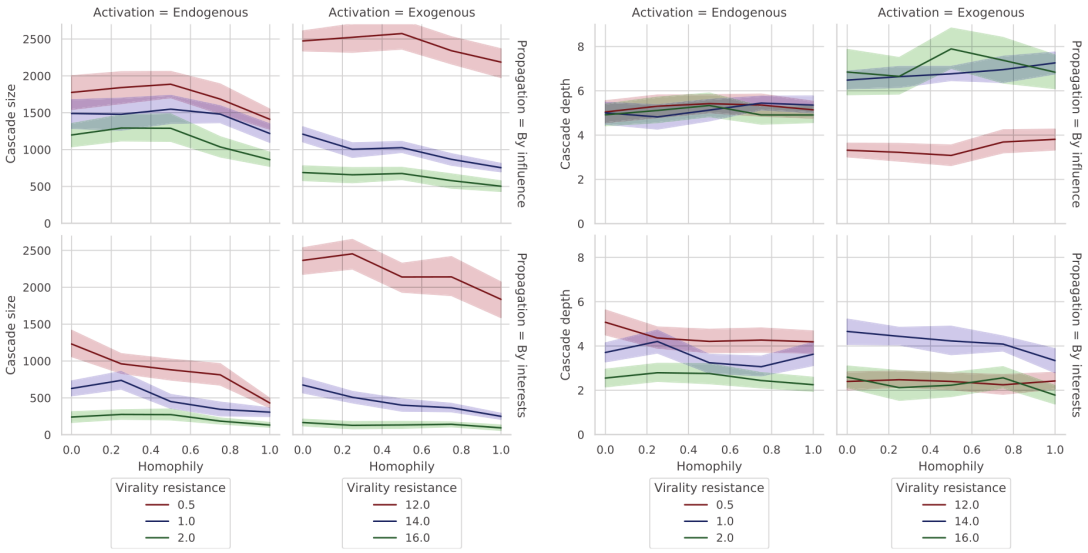


Fig. 2. For each of the configurations, the eventual cascade size (left) and cascade depth (right).

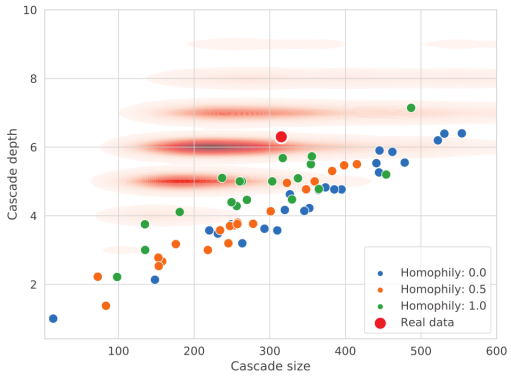


Fig. 3. Comparing real and generated cascades.

3.4.2 How do different configurations affect cascades. By tuning 4 parameters the authors create different configurations to start the information cascades. In this section we see the correlation between each of these parameters and how the cascade size and cascade depth changes as the parameter is tuned. The Fig. 2 shows results for each of the parameter.

(1) Propagation by influence results in cascades with bigger sizes and depth overall compared to just using propagation by interests. This clearly suggests that information propagated by influencers tend to reach more users and create more successful cascades.
(2) Exogenous activations tend to be more successful than endogenous activations. This suggests content spread to more users initially is more likely to cause bigger and even deeper cascades. Since endogenous content is initiated by a single user in the network, it is less likely to spread to more users. It is also worth noting the virality resistance is increased when spreading content exogenously.
(3) Surprisingly in many combinations of configurations, having a high degree of homophily results in tall cascades - deeper and smaller, while lower homophile creates flat cascades. The reason for this might be that high homophily can create echo chambers cliques that gets exhausted much faster and the propagation stops as it reaches the echo chamber's boundary.
(4) Having high virality resistance creates smaller cascades in terms of their sizes, however in the case of having an exogenous initial activation, high virality creates deepest cascades (although smallest in size).
(5) The most successful cascade with the largest size and depth can be achieved by having exogenous initial activation, propagation by influence and 0.5 homophily. This configuration also creates the deepest cascades. The most unsuccessful cascades (smallest and least deep) are generated by having exogenous initial activation, propagation by interests, 1.0 homophily and a high resistance to virality. This is because having propagation only by interests, the initial activation will activate all interested users, having a high homophily means all users are connected closely to these activated users and therefore in only a few iterations all interested users are exhausted.

4 CAN CASCADES BE PREDICTED?

4.1 Introduction

The authors show whether its feasible to predict whether a cascade in its initial state grow bigger. Using images posted on Facebook, they identify small cascades and create a deep learning model trained on features from 4 different perspectives: 1) **User**, 2) **Content**, 3) **Structure**, and 4) **Temporal**. After training the model they use explainable machine learning techniques to understand which perspective provides the most information to predict future stages of a cascades and consequently and how they impact the information cascade. Their prediction model show promising results in predicting the success of cascades.

4.2 Problem Statement

Facebook is an online platform where users create their profiles and add friends and pages to their account. Users are able to post content such as text, images and videos on their profile, this posted content then shows up on their friends homepage. Similarly on their own homepage the users are able to see content shared and posted by their friends and the pages they follow. Users also have the ability to like, comment and more importantly to this study share any content posted by another user or page. Facebook can be represented as a graph where each node is a user or a page. Edges between users are bidirectional representing their friendship and how information can flow between them. Edges between a user and a page is unidirectional where the information only flows from a page to its followers. It is important to have understand the distinction between the content posted by a user and a page. Pages tend to have more followers and in their preliminary study the authors show that content posted by a page has more likelihood of becoming a cascade, since in their dataset 81% of the cascades are initiated by a page with 11% of them reaching 100 reshares.

Cascades on Facebook. In their preliminary analysis on cascades on Facebook they show large cascades (of size bigger than 10) are extremely rare and even for cascades with identical images very few cascades grow big in size. This shows that the data set for cascades is extremely skewed towards them being unsuccessful and baseline results will show >90% accuracy. To solve this issue the authors redefine their problem, rather than predicting whether a cascade will be successful or not they predict the next stage of the cascade. Their formulation of the problem takes a cascade of size k and predict whether it will grow bigger than f(k). f(k) is the median size cascades of size k typically reach. They show in their results that this median size follows a pattern and for a cascade of size k the median is 2k, this is true for any k. Now they design a binary classifier that predicts whether cascade of size k will be greater than 2k or not.

4.2.1 Research Questions. In this section we list the core questions the authors try to answer with their study.

(1) **Can cascade growth be predicted?** The authors start of with answering whether cascade growth can be predicted or not. By showing that cascade growth is indeed predictable the authors show that features from an initial stage of a cascade can expose information about the future of a cascade. These features, since part of the initial cascade stage, if changed will result in different eventual cascade. Therefore, these features related to the initial structure, users, rate of growth and content do impact the cascade size at the end.

(2) **What are the factors driving cascade growth?** The authors then look at the factors and features that drive cascade growth. Their goal in this task is to look at the impact each feature has on the eventual size of a cascade. Furthermore, they also look at in what ways do these factors affect cascade growth structurally as well.
(3) **How predictability changes with observation window?** In their analysis of predictability of cascades the authors use initial stages of a cascade to predict the future of cascades. In this question, they try to answer whether increasing the observation size makes this prediction easier or more difficult. They also look at how the feature importance changes as the size of the observation window changes.

4.3 Methodology

4.3.1 Can cascade growth be predicted.

Predicting the next stage of a cascade. The problem statement for this part of the study is to predict the next stage of a cascade rather then predicting will a cascade grow bigger. Although the two problems might seem similar, as we discussed before predicting will a cascade grow is a more difficult problem due to the skewness of the data. Therefore, the authors formulate the problem in a different way, given a cascade of size k will the cascade grow bigger than f(k) the median eventual size of cascades of size k. Since the they are seeing whether the size will grow more than the median size the skewness by definition becomes 50%.

Dataset description. To train their binary classifier the authors create a dataset of photos uploaded on Facebook on the month of June 2013. They collect all publicly available images that have at least 5 reshares (i.e k=5). They also have the whole network structure made available to them by Facebook along with information about each user and each post made. The total number of cascades in their dataset that have k=5 are 150,572.

Training the model on k=5. The authors dataset consists only of cascades that have are of at least size 5. They create a data set of cascades at their point in life when their size was 5. The median eventual size (f(k)) of all these cascades is 10, therefore, the prediction task is to predict whether these cascade will get bigger than 10 or not. With the observation window of a cascade limited to k=5 the authors extract all the features mentioned in the following section. The authors perform correlation analysis on each of the feature as well, the results of which are explored in the following sections. The authors train different varieties of machine learning models from SVM, naive Bayes, decision trees, random forests, linear regression, and logistic regression. They report their results of the logistic regression model due to he simplicity of explaining the importance of each features.

4.3.2 What are the factors driving cascade growth.

Explainable Machine Learning. In this section we describe how they authors try to answer this research question. By creating a predictive model trained on cascades on Facebook and their features the authors are able to predict the future stages of cascades and whether they will grow or not. Predicting the growth of a cascade based on features suggests that these features, in the initial stages of the cascades, play an important role in determining the future o the cascade and essentially drive the cascade growth. Given a set of cascades in their initial stages and a label on whether they will double in size, the model is

able to make relationships between which features are most correlated with the growth of the cascade. Furthermore, the authors also look at the simple correlation coefficient between each of the feature and the cascade growth.

Features. As mentioned before the authors train their model on 4 different sets of features. We explain each of the 4 feature set in this section.

(1) **Content** This feature looks at how the features of the content contribute in whether a cascade will double or not. Understanding the importance of this set of features will show how much content plays a role in its virality and why cascades with identical content have different results. To collect these features, the authors use a linear SVM model to extract features from the image. Furthermore, they also extract features form the accompanying caption.
(2) **User** The root of the cascade is an important predictor of whether a cascade will succeed or not. Many previous works studying information cascades on Twitter have shown the importance of a well connected root. The authors extract features related to the user who started the cascade such as their popularity, gender and activity. Additionally, since in this study the model is given a cascade of size k, we can extract user feature for the other k-1 resharers as well. This helps the model understand the demographic of the users already activated on the contagion.
(3) **Structure** These features captures the network of the existing cascade of size k. This helps the model understand how the information has spread so far. The authors also provide potential graph as well which is the network graph of the k sharers regardless of whether the nodes have shared the contagion or not. This enable the model to create a representation of the potential spread of the contagion and make better predictions for the growth of the cascade.
(4) **Temporal** These features captures the rate at which the current cascade has grown since its inception. Since the speed at which the contagion is being reshared changes, these features also capture the rate at which the speed of the cascade growth is increasing or decreasing. Since it is more likely that a dying cascade will have a slower rate of reshares.

Controlling for content. The authors look at the importance of the content perspective specifically. In this experiment the authors control for content and observe how same content with different initial configurations can result in different eventual cascades. They sample cascades from Facebook that have identical content and look how the cascades grow differently. They also observe how other features impact the predictability of their model in these cases.

4.3.3 How predictability changes with observation window size.

Predictability and the observation window. In this part of the study he authors try to understand how the accuracy of the model improves as the model is exposed to more information by increasing the observation of the cascade i.e. increasing the k. Increasing the amount of information might help in improving the accuracy but it is important to understand that increasing k would in turn increase f(k) as well. This would mean that the model would have to predict the evolution of the cascade more into the future. This study also opens an opportunity to understand how the importance of the feature changes as the cascade size grows bigger. For this analysis the authors design an experiment where they have a minimum reshares of at least R. The objective of the model is to predict whether the model will grow bigger than R. The authors modify the size of the observation window and see how the performance of

the model changes. The observation window is the size of cascade at which the authors extract the features and train the model on. The authors start with a value of k and keep increasing it to see the effect of observation window size on the performance of the model and also how the impact/importance of each of the feature changes as the model is trained on cascades in later stages rather than initial.

4.4 Results

4.4.1 Predictability of cascades. Using the cascades of k=5 the authors train their binary classifiers on all of the aforementioned features. Reporting the results for the logistic regression they show an accuracy of 0.795 and AUC of 0.877. These results prove the predictability of cascades and the hypothesis that the features used to train the model are variables that do affect the propagation of information. Next, using only temporal features to train the model they show that these are the most important among all other features in the predictability of the cascade's future. Furthermore, by limiting different feature sets (such as training on all features except temporal) they show in Fig 4 the robustness of the model by achieving similar accuracy scores.

4.4.2 Factors driving cascade growth. Using correlation coefficients the authors understand the importance of each of the feature towards predicting the eventual size of a cascade. Consequently, they are able to infer which features are important in driving the cascade growth. In Fig. 6 they show the importance of each feature and how it changes over time.

Cascades with identical content. In this experiment the authors show in a preliminary analysis how there are many cascades with identical content but drastically different eventual cascade size. By controlling for content when training their model they observe how user features, structural features, and temporal features play a much important role in predicting the cascade eventual size. In conclusion they show content is not the driving force for a cascade and its lack of virality can be overcome by manipulating other initial configurations such as the root user and giving the content a temporal boost.

4.4.3 Predictability and observation window. By having the ability to change the observation window, k, the authors are able to show how the predictability of cascades future changes as the cascades grows larger in size and the model has to predict more into the future. Using different values for k (5 and 25), they show that the model does improve in accuracy and is able to make better predictions. They show in Fig 5 that as the observation window increases in size the model improves linearly. This shows that the addition of a single observation improves the predictability for a cascade of size 5 and a cascade of size 100 in the same magnitude. This linear increase in the performance also shows that more information is always better and there are no diminishing returns.

Changes in feature importance. By varying the observation window available during training the model the authors see how the importance of features changes. Here the authors show which features are more important in the initial stages of a cascade and which are important in the later stages of the cascade. The authors show as the cascades grow the importance of the root node and content decreases. This is especially true for content features (such as did the image have a caption, was it in English and whether an image was a meme) which decrease in importance to 0 for bigger values of k. Regarding the structural features, connectedness of nodes is important in the beginning but as the cascade grows the importance starts to decrease. Finally, for temporal features, predicting for bigger values of k, the

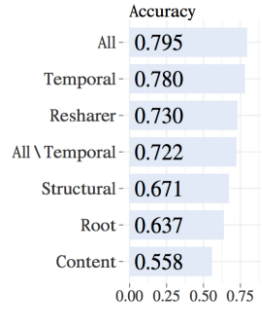


Fig. 4. Feature importance based on model accuracy.

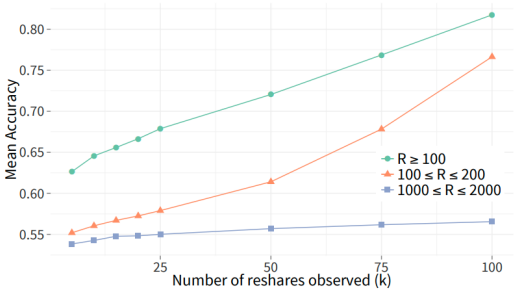


Fig. 5. Model performance as the observation window is increased.

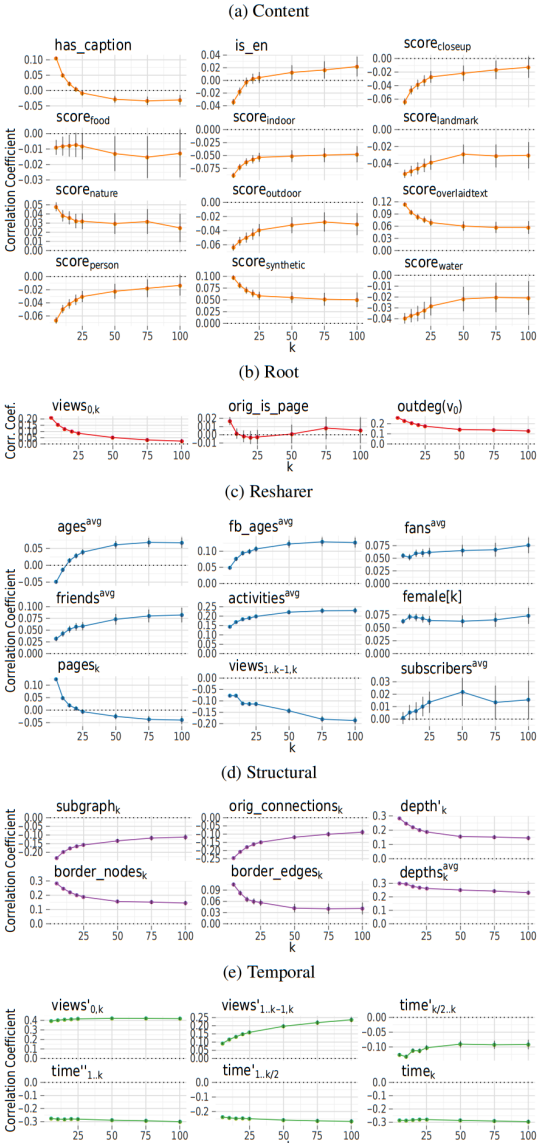


Fig. 6. Importance of each feature as the observation window k is increased, shown by correlation coefficients

early amount of views and conversion rate. Other temporal features remaining consistently important regardless of the cascade size. Furthermore, the authors show how the correlation coefficient changes for each of the feature as the observation window increases for the model.

Predicting Cascade Structure. Similar to how important understanding how cascade size changes with each variable, examining eventual cascade structure under different initial configurations can aid in

maneuvering how the cascade evolved and which groups of people it reaches. Cascade structure can vary from being tall cascades i.e. long and narrow, flat cascades i.e. wide and short, and anything in between. The authors look at the same parameters as before to predict cascade structure given its initial stage and in turn understand the importance of each of the feature and how it impacts the cascade structure. The authors show how in predicting the structure of the cascade, temporal and structural features are still the most important. Unsurprisingly, they observe structural features becoming more important in predicting the final structure as the observation window is increased.

5 CONCLUSION

In this report we look at three papers that try to understand the dynamics of information flow using different perspective and different methodologies. The first paper look at the dynamics of information flow by performing an empirical study on existing platforms. They look at how different platforms react differently to cascades. Furthermore, using Hawkes process the authors look at how each platform influence each other. Using this empirical method they show how content and platform features can impact the size and structure of cascades. Their methodology is however lacking in not being able to make direct quantative correlation between the differences each platform has and the different cascade that result because of these differences. The second paper creates synthetic cascades to understand how different initial configurations result in different cascade structure and sizes. They have the ability to tune the parameters to generate any desired cascade. This work overcomes the shortcomings of the first paper by making direct quantative correlation between the parameters their model can tune and how the resulting cascade grows. However, since the cascade they generate are synthetic and they are unable to model every single variable involved in a real world cascade. Finally, the third paper tries to predict eventual cascade sizes and use explainable machine learning techniques to understand how each feature impact the cascade size differently. Using real world cascade from Facebook they extract features from their initial stages and predict whether the cascade will grow or not. Using these set of features they are able to predict whether the cascade will grow or not. Therefore, quantify how much the features drive the cascade growth.

REFERENCES

[1] 2013. *Reddit's 'Find Boston Bombers' Founder Says 'It Was a Disaster' but 'Incredible'.* https://www.businessinsider.com/reddit-falsely-accuses-sunil-tripathi-of-boston-bombing-2013-7

[2] 2014. *Dissecting the PizzaGate Conspiracy Theories.* https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html?searchResultPosition=1

[3] 2014. *The only guide to Gamergate you will ever need to read.* https://www.washingtonpost.com/news/the-intersect/wp/2014/10/14/the-only-guide-to-gamergate-you-will-ever-need-to-read/

[4] 2019. *8chan, 8kun, 4chan, Endchan: What you need to know.* https://www.cnet.com/news/8chan-8kun-4chan-endchan-what-you-need-to-know-internet-forums/

[5] 2019. *Top Twitter Demographics That Matter to Social Media Marketers.* https://blog.hootsuite.com/twitter-demographics/

[6] 2020. *Nearly half of Twitter accounts pushing to reopen America might be bots.* https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/

[7] 2020. *What Is QAnon, the Viral Pro-Trump Conspiracy Theory?* https://www.nytimes.com/article/what-is-qanon.html

[8] Yochai Benkler, Rob Faris, and Hal Roberts. 2018. *Network propaganda : manipulation, disinformation, and radicalization in American politics.* Oxford University Press.

[9] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web.* 925–936.

[10] Federico Cinus, Francesco Bonchi, Corrado Monti, and André Panisson. 2020. Generating Realistic Interest-Driven Information Cascades. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 107–118. https://www.aaai.org

/ojs/index.php/ICWSM/article/view/7283
[11] Yosh Halberstam and Brian Knight. 2016. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* 143 (2016), 73–88.
[12] Young-Jin Lee, Kartik Hosanagar, and Yong Tan. 2015. Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Science* 61, 9 (2015), 2241–2258.
[13] Qihua Liu, Shan Huang, and Liyi Zhang. 2016. The influence of information cascades on online purchase behaviors of search and experience products. *Electronic Commerce Research* 16, 4 (2016), 553–580.
[14] Qihua Liu, Xiaoyu Zhang, Liyi Zhang, and Yang Zhao. 2019. The interaction effects of information cascades, word of mouth and recommendation systems on online reading behavior: An empirical investigation. *Electronic Commerce Research* 19, 3 (2019), 521–547.
[15] Efrat Nechushtai and Seth C Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (2019), 298–307.
[16] Richard A Stein. 2017. The golden age of anti-vaccine conspiracies. *Germs* 7, 4 (2017), 168.
[17] Nathalie Van Raemdonck. 2019. The Echo Chamber of Anti-Vaccination Conspiracies: Mechanisms of Radicalization on Facebook and Reddit. *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming* (2019).
[18] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*. 405–417.