Cascades on Online Social Networks: A Chronological Account

Nora Alrajebah, Thanassis Tiropanis, and Leslie Carr

University of Southampton, Web and Internet Science, Southampton, UK
N.Alrajebah@soton.ac.uk,tt2@soton.ac.uk,lac@soton.ac.uk

Abstract. Online social network platforms have served as a substantial venue for research, offering a plethora of data that can be analysed to cultivate insights about the way humans behave and interact within the virtual borders of these platforms. In addition to generating content, these platforms provide the means to spread content via built-in functionalities. The traces of the spreading content and the individuals' incentives behind such behaviour are all parts of a phenomenon known as information diffusion. This phenomenon has been extensively studied in the literature from different perspectives, one of which is cascades: the traces of the spreading content. These traces form structures that link users to each other, where these links represent the direction of information flow between the users. In fact, cascades have served as an artefact to study the information diffusion processes on online social networks. In this paper, we present a survey of cascades; we consider their definitions and significance. We then look into their topology and what information is used to construct them and how the type of content and the platform can consequently affect cascades' networks. Additionally, we present a survey of the structural and temporal features of cascades; we categorise them, define them and explain their significance, as these features serve as quantifiers to understand and overcome the complex nature of cascades.

Keywords: Social Network Analysis, Information Diffusion, Cascades

1 Introduction

Since its emergence, the Internet has created a venue for human-to-human social interaction. In fact, the demand for some form of social networking was raised early on. This was facilitated by different types of computer-mediated communication (CMC), where 'humans' communicated with each other via the "instrumentality of computers" [34]. This comes in many forms on the Internet such as instant messaging, emails, and chat rooms. CMC was the focus of much research in the effects of such communication on social systems. In fact, Kaplan and Haenlein [36] stated that " ... the Internet started out as nothing more than a giant Bulletin Board System (BBS) that allowed users to exchange software, data, messages, and news with each other." This statement emphasises the core

purpose of the Internet as a medium that facilitates different forms of social interactions.

The invention of the Web in 1989 added another dimension to communication on the Internet, providing a wide range of possibilities for human interaction [11, 1]. The advance of the Web 2.0 offered a variety of applications that fundamentally changed the way users communicate such as wikis, blogs, RSS, podcasting, and social networks [39]. Therefore, in addition to communication and collaboration, individuals began to contribute to the Web by adding user-generated content. That is what differentiates Web 2.0 from the previous Web [49]. Online social networks have seen a popularity surge following the proliferation of Web 2.0 applications [33]. However, their core purpose is not new; they merely emphasise the Internet's main purpose: facilitating the exchange of information between its users [36]. The Web offers an enormous amount of data that can be analysed to cultivate insights about the way humans behave and interact with each other online. Simultaneously, with the advancements of the Web technologies, a phenomenon has been observed; in the early days on blogosphere, bloggers would share the same URL after being exposed to it by other bloggers, in a cascading manner that can be traced. This phenomenon is information diffusion, and it is concerned with studying the way information is spread on the Web. Online social networks have proven, in many occasions, their vitality for a range of activities that are powered by information diffusion, such as: mass convergence and emergency events [35], spreading information about good practices such as saving energy on earth day [20], bringing people's attention to incidents that might lead to 'public shaming' behaviour [45].

There are three components in any information diffusion process: the content, the context that facilitate the diffusion and the outcome of the process which is the cascade [5]. The earliest research in this field studied diffusion in the blogosphere [4, 41]; as new platforms have emerged, they have been used to analyse information diffusion dynamics [38, 24, 3, 40, 10]. Research in the field of information diffusion varied according to the purpose of study and the diffusion component(s) that is been taken into consideration. Hence, in their survey of information diffusion in online social networks; Guille et al. [32] categorised the research challenges and approaches in the field into three categories: 1) detecting popular content, 2) modelling information diffusion, and 3) Identifying Influentials.

In this paper, our focus is on cascades, which are defined as the structural representation of the information diffusion and are often perceived as the final outcome of the process [42, 25]. Cascades are amplified on online social networks platforms by built-in mechanisms that allow users to share content while crediting the source or the person who posted it [15]. The aim of this paper is to provide a holistic overview of cascades based on their relatively long research history; exploring how cascade-related research has evolved throughout the years. We will attempt to answer the following questions about cascades: *(a) What are they? (b) What is their significance?, (c) How cascade networks are constructed? (d) What are the effects of the content type and the available data on cascade*

construction approaches?. (e) What features we can use to analyse cascades? i.e., how can we quantify cascades?

The first two questions set the scene; they emphasise the importance of cascades analysis as a proxy to unveil the sharing dynamics between users on the Web. The answers of questions c, d and e will be presented as a review of two aspects related to cascades: cascade networks construction and cascade features. We will discuss how these two aspects change depending on the content type in the diffusion event and the platform's functionalities. In addition, we will look into the impact of the data that is made available for collection on cascade's analysis. By tracing cascade-related research across platforms, we aim to provide an overview of cascades, which will help designing research problems, and will help researchers as guidelines to construct and analyse cascades.

This paper is organised as follows, Section 2 provides some background including cascades definitions, significance and purposes. In Section 3, we will look into cascade construction and the different construction approaches. Section 4 is dedicated to cascade features, it is divided into two parts: the structural features and the temporal features. And finally Section 5 concludes with some remarks about cascades and how this paper can help researchers who would like to study them.

2 Background

2.1 What Are Cascades Networks?

Networks, in their general sense, are structures that consist of a set of nodes and links; the links associate nodes with each other, encapsulating a specific type of a relationship between the two. In mathematical terms, networks are modelled as graphs with vertices and edges [48]. The core concept of networks is their 'connectedness', a phenomenon that has been observed in fields such as Biology, Computer Science and Sociology, and it arises from the flexibility of the definition [22]. A social network can be defined as a network where the nodes represent people and the links represent the relationships and interactions between them [37, 48]. Examples of such relationships are: acquaintance, friendship, co-authors, co-workers, affiliation, family relationships, information exchange, etc. [29]. All of these networks link people and, via these links, people interact with each other for many purposes such as: talking to each other, sharing information, and collaborating. One example of such networks are cascade networks, which are networks that link people based on the direction of the flow of information/content between them.

For economists, information cascade occurs when an individual decides that it is optimal to follow the behaviour of those before him after observing their behaviour, without taking into account his own information [13]. In fact, economists differentiate between information cascade and herding behaviour. The difference between the two is that in information cascade individuals decide by making inferences ignoring their own information, while in herding individuals follow

the 'herd' without necessarily ignoring their own information [16]. Nevertheless, the term 'cascade' was picked by researchers to describe a similar phenomenon that has been observed in OSNs. In cascades, messages travel through the social network links from one user to another [38]. When gathered, the paths that these messages travel through create a network that resides as a layer on top of the social network. These networks are the cascade networks and the paths messages take are often called information paths [28].

A cascade as defined by Goel et al [25], comprises a seed individual who shares an item of information independently from any other individual, followed by other individuals who are influenced by the seed to share the same information. Another definition by Leskovec et al. [42] state that cascades are phenomena caused by individuals' influence in which an action or idea becomes widely adopted by others [26, 30]; hence, they are known as 'fads' [13]. Both definitions emphasis one point: cascade networks are structures that represent (and preserve) the relationships between users as they share the same content.

2.2 Significance of Cascades

Analysing cascades is an essential step towards understanding the way information propagated on the Internet [21]. When content spreads, it provides us with valuable information about the users involved in the process. As we mentioned in the previous section, cascades represent some form of a relationship between the users. This relationship has been identified in the literature as influence [21]. Identifying influencers has received a significant attention in previous work, and cascades were considered as indicators of influence. Hence, the paths that information takes to reach individuals are recognised as influence paths in many studies, as they directly indicate that one user influenced another to spread the message. In addition to influence, researchers identified another reason behind sharing the same content they do so because there is some degree of homophily among them. Dow et al. [21] state that a user's repeated exposure to a particular content increases the chances of sharing it. They argued that in such a case, these users are subject to both influence and homophily [10]; repeated exposure increases the influence factor, and being surrounded by a group of users who are susceptible to an item means that the user himself is susceptible too. However, cascades do not occur because of influence and homophily only, as both are tied to the nature of content as well. Hence, a cascade informs us about the value of the content itself. Given that users have limited attention, a successful cascade is the one that gets the most attention across the competing cascades at a particular moment [53, 47].

Bild et al. [14] refer to cascade networks as implicit networks because they are constructed using a subset of the social network, which they define as an explicit network. They argue that analysing cascade networks is important as these 'implicit' networks can serve as an accurate indication of interest or trust relationships. They conjecture that cascade networks model real-world social, interest and trust networks better than the social network. They argue that connections on the social network (follow/friend) entail that users are willing to

listen to each other, but connections on the cascade network are better indicators because they are created using a forceful sharing action that pushes the content to the users list of friends.

Furthermore, analysing cascades can help detect network evolution and link creation, since users often create new links (follow/friend new users) after being exposed to novel information sources. Myers and Leskovec [46] studied the relation between cascades and the creation of new links in the social network. They related the sudden bursts of connectivity to the dynamics of sharing on Twitter. Antoniades and Dovrolis [8] used the number of retweets and follow reciprocity to model link formation. They also studied link removal dynamics on Twitter after reading a tweet or receiving a retweet from the user. Also, Farajtabar et al. [23] introduced a model that takes into account both activities (sharing and link creation).

To summarise, cascades play an important role in different social network research endeavours:

1. They allow researchers to estimate influence and homophily between users.
2. They work as a proxy to estimate the value of the content that spreads, as successful cascades means that the content attracted a larger number of users.
3. They are better indicators of users' interest and trust relationships than the social network.
4. They help explaining social network evolution and link creation and removal.

2.3 Purposes for Studying Cascades

Cascade studies' purposes vary depending on the objective of the study and the data available for the researchers. Throughout the years, and the different platforms that have been investigated, research purposes have ranged from merely observing and quantifying cascades, to tracking them, predicting information flows, and modelling them [31].

Figure 1 illustrates the four general perspectives for studying cascades. The first, and the essential purpose, is tracking existing cascades then either constructing or inferring them. The ability to construct a cascade depends solely on the data available during the construction process. We will look into this in details in the next section.

The second perspective focuses on quantifying cascades, structurally [21], temporally [31], or just numerically, combined with some platform dependent measures [10]. Often, cascades' tracking is the initial step before quantifying them. For instance, the structural analysis of cascades requires constructing cascades first before the analysis phase can take place. However, some studies focuses on analysing cascades quantitatively, thus, they do not attempt to construct cascade networks as the structure of cascades is not essential for this purpose, e.g. [10]. In Section 4 we will present a survey of the structural and temporal features of cascades. We will highlight their significance to understand cascades.

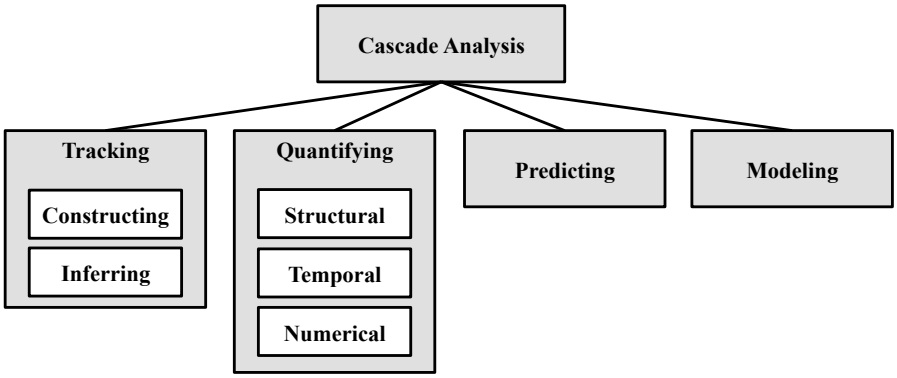


Fig. 1: Purposes for studying cascades

The third perspective looks at modelling cascades, i.e. using generative algorithms to create cascade networks using the characteristics observed from the tracked cascade networks [31, 43]. The fourth perspective investigates predictions such as the likelihood that a piece of information will be shared in the first place [50], or the possibility that a popular piece of content will continue to be popular [44], or predicting the future growth of a cascade [18]. Most of the time, one study incorporates one or more purposes in its analysis. However, this paper is focused on the first and second purposes, namely: tracking and quantifying cascades. The third and fourth perspectives are beyond the scope of this paper, thus, we briefly mentioned them here to provide an overall view of the purposes of studying cascades.

3 Constructing a 'Cascade' Network

As we mentioned earlier, within social networks, many sub-networks can be created using the same nodes that can be linked using edges with various meanings. As soon as information starts to spread within a population, another layer could be added on top of the original network to represent the flow of information [28]. This is often called a diffusion/propagation network or a cascade network [22]. Using Twitter as an example, instead of creating a network of followers, we could create a network where each node represents a user and each link represents a retweet direction. Thus, if A retweeted a tweet posted by B, then there would be a link from B to A, creating what is known as a 'retweet network' [56], or as we will refer to it here a 'cascade network'.

As we mentioned in the previous section, to track existing cascades, they must be either constructed or inferred. In the early studies of cascades, in blogs, for instance, there were no built-in mechanisms for diffusion; thus, most of the early studies used various features to infer cascade networks. Adar and Adamic [4] added a link between two blogs if there is an explicit link to the other. If there

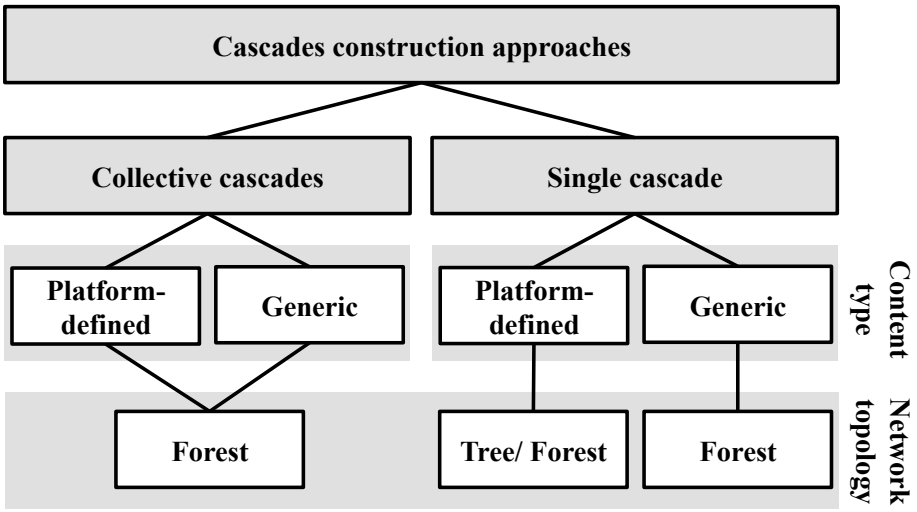


Fig. 2: Cascade networks construction approaches and their resulted topologies

is no explicit link, they infer it using a number of features related to the blog network structure, historical data about the blogs' posts, text similarity, and timestamps. Most early studies of cascades on online social networks exploited users' written credit attribution of content sources to infer cascade networks. Examples of credit attributions are "RT", "via", "retweet", and "reshare" [21]. There were also many attempts to use the social network and timestamps to infer cascade networks [27]. However, with more contextual information available it is possible to construct more accurate cascade networks. For instance, Dow et al. [21] used information about reshares, timestamps and clicks on feed, to infer cascade networks and compare them with cascade networks constructed solely from tracked information.

More recently, online social network platforms start incorporating the ability to share content with a click on a button; for example retweet on Twitter, Reblog on Tumblr and Share on Facebook. With these functionalities in place, users can share different types of content easily. As a consequence, tracking existing cascades is now feasible with the appropriate access to data. Thus, researchers are now able to construct existing cascades directly from the platform.

3.1 Cascade Networks Topologies

A cascade is often perceived as a tree that has a single root (the cascade initiator) which is linked to other nodes. Further nodes can be added by linking to the existing nodes in the cascade network and all of the added links follow a strict time order [41]. However, cascades are not always shaped as trees, in fact, their structure changes depending on the type of content these cas-

cades networks represent. Anderson et al. [7] classified cascade networks into: information-sharing networks in which information spread between the users and signups which mimic the adoption of a new technology. This classification does not specify the topology of the generated cascade network. Thus, here we present a different classification of cascade networks based on their topology. The basis of this classification is the content type and the diffusion mechanism provided by the platform.

There are two main approaches to constructing cascade networks that have been used in research. Figure 2 illustrates them and the resulting cascades' topologies generated by each approach. The first approach is collective cascades, in which a large cascade network is constructed, linking users according to their sharing activities (retweet/reblog) collectively for a group of cascading items. The topology of this network is a forest that has several components. These large networks are useful to study the sharing activity patterns within a platform [54, 14]. Collective cascade networks are often weighted to represent how often a link occurs between two nodes [41].

The second approach is for single cascades in which cascade networks are constructed for each item that has been shared separately. Of the two categories of content, the first is a platform-defined elements such as a tweet in Twitter or a post in Tumblr. The second category (generic elements) covers any element that can be embedded within platform-defined elements such as a URL, a hashtag, a text, or a photo. Different content types require different data collection and analysis methods, and they create a completely different network topology.

The platform-defined elements that can be shared are for example: a post on Tumblr and Facebook, or a tweet on Twitter. This type of content spreads via explicit diffusion functionalities such as retweeting, sharing or reblogging. Their spread generates cascades that can be tracked or inferred on the platform. Cascades are constructed from the flow of information from users who might or might not be connected to each other by a relationship within the social graph [18, 5, 19, 6]. These cascade networks ideally follow a tree topology; the root is the source (author) and from there content travels across the social network. However, in many cases due to the limited access to the platform, some data might be missing because it is deleted, the topology of the generated cascade network will be a forest where there will be separate components for each isolated part that can not be linked to the main tree due to missing data [52, 6].

Because the diffusion of generic elements, such as hashtags and URLs, does not occur via explicit diffusion functionalities in social networks. Thus, timestamps are often utilised as an indicators of diffusion between users assuming that these users have an established social relationship in the social network graph. Cascade networks of generic items are different to cascade networks of one story. These networks incorporate multiple introductions of the same item in the network, thus naturally their topology will be a forest with separate components (sub-cascades). Hence, the number of sub-cascades and their sizes can be used as structural features of these networks [24].

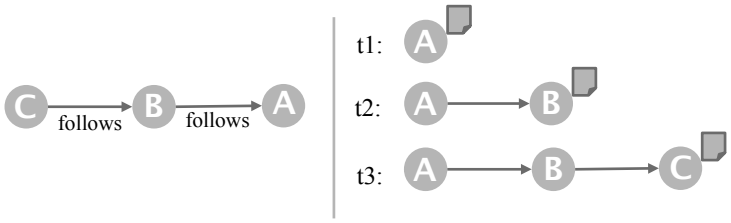


Fig. 3: Link types in cascade networks, left:relationship perspective, right: information flow perspective

Collective cascades networks can be easily converted into single cascade networks by separating the different branches of the network where they are related to the same story (message). For instance, Leskovec et al. [41], generated cascade networks following the two approaches from blogosphere. They constructed a post network that links posts if they credit each other. From the post network they constructed a blog network by collapsing the links between blogs and assigning weights to them. Following this method, they constructed separate cascade trees from the post network. Sections 3.3 and 3.4 will discuss the cascade construction approaches used in different platforms, including the data used for their construction, the detected diffusion mechanism, and the structure of the cascade networks.

3.2 Link Direction

Edges between the nodes in a network might convey different meanings. For instance, Bild et al. [14] identify the number of users who retweet from a user as the popularity; while the prolificity refers to the number of users a user retweet from. Hence, the direction of edges in a network can have different meanings. Consequently, all the measures that rely on the edges' direction will be affected.

Figure 3 illustrates two possible uses for edges' direction as used in the literature For example, suppose that we have three users, A, B and C. For simplicity, suppose that we have the following settings: user B follows user A and user C follows user B. Then, each time user A posted some content user B will be exposed to it and when user B shares that content after seeing it; user C will be exposed to the content too and can share it as well. In such a scenario, there are two possible representations:

1. **Relationship perspective**: If our concern is to represent who is linked to whom i.e., who follows whom, then the in-link from B to A means that B is linked to A, and the in-link from C to B means that C is linked to B. This is shown on the left in figure 3, this representation is often referred to as the social network or the follow network.
2. **Information flow perspective**: In this case, the in-link from one user (A) to another (B) means that B is exposed to whatever information A

has and when B shares that information too an edge will be drawn from A to B indicating the flow of information from A to B. This representation is often used for cascade networks. Figure 3 shows how this network can be constructed cumulatively at different timestamps. At timestamp t1, A posted a content, then when B was exposed to it, B decides to share it at timestamp t2, hence the edge from A to B and so on.

3.3 Cascades in Blogs, Recommendation Networks and Internet-chain letters

Data used to construct/infer cascades:
As mentioned earlier, in the early days of blogosphere there were no convenient mechanism to share content. Thus, instead of following the traces, cascades are inferred using a variety of measures such as: posts text, explicit links to other blogs, features about the blogs network, the blog and the timestamps [4]. In another study of cascades on blogs, the In-links/out-links between blog posts and timestamps were utilised to construct cascade networks [41].

On the other hand, on recommendation networks information about: products, time of recommendation, whether the product is purchased, and time of purchase are utilised to infer these networks [42]. Also, Liben-Nowell and Kleinberg [43] used the ordered list of users who forwarded the petition to construct the cascades of chain-letters.

Diffusion mechanism:
As we can see the lack of explicit diffusion functionality means that various mechanisms of diffusion were identified such as: posting a URL in a blog [4], recommending a product [42], linking between posts on blogs [41], and forwarding of a petition letter from one user to another [43].

Cascade networks topology and components:
The network topology of these cascades and their components vary based on the platform and the purpose of them. For instance, in [4] the cascade networks structure is trees, where the nodes are blogs and the edges between them are inferred to show the direction of diffusion of information between the blogs. while Leskovec et al. [42], constructed a posts network that links posts in different blogs, and a blogs network which is a collapsed and weighted version of the posts network. Both networks are forests and they extracted separate cascade trees from the posts network. On recommendation networks a separate group networks and a product networks are constructed, where the nodes are the customers and the edges connect customers' product recommendations [41]. Finally, in the work Liben-Nowell and Kleinberg [43] the lists of users in each petition contains duplicates or missing users. Thus, the cascade networks are trees inferred by removing edges that did not appear in a sufficient number of copies. Thus, the nodes are users and the edges represent the direction of information flow between them .

3.4 Cascades in OSNs

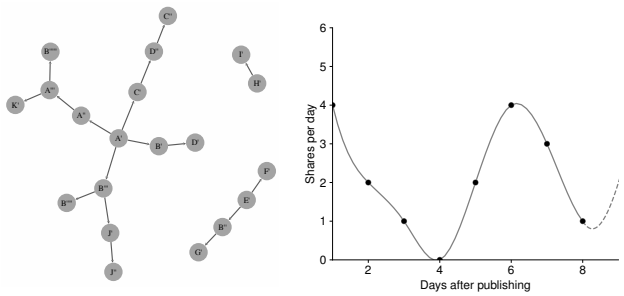
Data used to construct/infer cascades:
Depending on the content type in each study and the diffusion mechanism, the data needed to construct cascade networks on OSNs vary from: retweets on Twitter [38, 12, 14], reblogs on Tumblr [17, 54, 5, 6], share on Facebook [21, 18, 19]. The tweet texts, timestamps and social network are used in [24] to infer cascade networks of URLs. Yang and Counts [55] analysed tweets' texts that contain topics and mentions of other users to construct cascades. Also, text analysis (status updates that include the meme and the words 'copy', 'paste' and 'repost'), lists of users who commented on users' status and timestamps are used in [3] to construct cascades of memes on Facebook. In another study of cascades of memes on Facebook, the social network, time, text similarity measures are used [2]. On LinkedIn signups and timestamp are used to construct cascade networks of invitations [7]. These examples shows the diverse views of cascades on OSNs; they show us how the diffused content type affects the cascade, and the varieties of data that can be used to either construct or infer cascade networks.

Diffusion mechanism:
On OSNs the main diffusion mechanism is provided by a platform's functionality (retweet, reblog, share). Other mechanisms of diffusion are: posting a URL [24], or crediting the source using 'RT @' in tweet text [24, 14]. For memes, the diffusion mechanism is simply copy and paste of textual memes [3, 2].

Cascade networks topology and components:
Various cascade networks topologies are used based on the content type, as mentioned earlier platform-defined elements generate trees, while generic elements generate forests. For example, Kwak et al. [38] created retweet trees for each tweet in their dataset and forests for each topic. Also, in [24], because the diffusion mechanism used is either posting a URL or crediting the source, the generated cascades' structure is a forest. Due to their nature, cascades of memes are forests [3, 2]. There are also two studies that constructed large cascade networks of collective cascades [54, 14]. In general, the nodes in most of the cascades on OSNs are users, and the edges always indicate the direction of information flow between them. An exception was found in [2], where the nodes are meme variants and the edges between them link a meme variant to its parent.

4 Cascades Features

In general, the data we can harvest about cascades is multidimensional in its nature. It has a twofold purpose: the first is to allow cascade networks to be constructed using the detailed information about users sharing from other users; the second is to allow the creation of a time series dataset, where the number of sharing activities at a given time (day or hour) after publishing is recorded. Figure 4 illustrates these two data representations that are used for the analysis. The first is linked to the relation between the users involved in the cascade, i.e.



(a) (b)

Fig. 4: Cascade data dimensions, (a): Cascade network data, (b): Time-series data

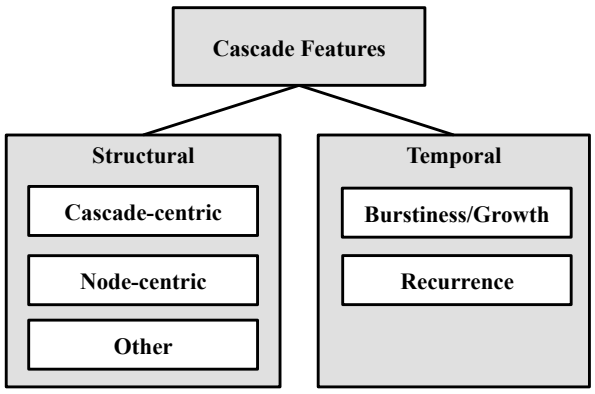


Fig. 5: Cascade features classification

who influenced whom to spread the content. The time-series information about cascades provides the number of diffusion events that occur at a given time. Each of these dimensions is related to a different aspect: either the structural or the temporal. These two aspects complement each other and provide a better understanding of cascades, as Scott [51] argued that the temporal aspect adds value to the structural aspect when analysing data from social networks.

The level of access researchers have to the platform's data determines the type of data they can gather. For instance, utilising a privileged access ensures that both dimensions are harvested, minimising the effect of missing or deleted data. In addition, with privileged access researchers have unlimited access to rich metadata such as clicks in News Feed [21]. As a result, they can infer cascades more accurately. Figure 5 illustrates the two classes of cascades features; in the next subsections, we will explore the structural and temporal features of cas-

Table 1: Structural Features

Feature	**Definition**
Cascade-centric features	
Depth, Range, distance to the root, Maximum depth, maximum hop count, cascade height [43, 24, 38, 9, 55, 3, 25, 21, 17, 6]	Represents the height of a cascade, it is calculated using the number of subsequent occurrences of message passing events, i.e. maximum number of hops or range of influence. Maximum depth and average depth can be measured too.
Width [43]	The maximum size of a set of nodes which share the same depth.
The fraction of nodes with exactly one child [43, 6]	How many nodes in a cascade with exactly one child.
Scale [55, 6]	The number of nodes influenced by a message at depth equals one.
Wiener index [18, 7, 5]	It is used to measure the structural virality of a cascade. It is computed as the average distance between all pairs of nodes in a cascade.
The percentage of adoption per depth [25, 7, 6]	The percentage of adoption events that occur at each depth from the root.
Number of nodes at depth = 1 [21, 6]	The number of nodes that are one step away from the author.
Connectivity Rate [52]	The percentage of users who have one edge at least, hence they were influenced by other users.
Root Fragment Rate [52]	The percentage of nodes that have either direct or indirect connection to the root node.
Diameter [42, 52]	The diameter of a network.
Node-centric features	
Fanout, Branching factor [31, 21, 5, 6]	The number of subsequent cascades that follow directly from a particular node (user).
Size of Sub-cascade [24, 21, 6]	The size of the sub-cascade under a particular node.
Other	
Frequency of distinct cascade structures [42, 41, 25, 17]	After constructing all of the cascade trees in a dataset, it is possible to compute the frequency of cascade structures. This process is computationally expensive as it aggregates all the generated cascade trees to identify similar structures, e.g. trees with root only, or trees with a root and two child nodes.

cades. We will identify these features and highlight their significance in relation to cascades' analysis.

4.1 Structural Features

Analysing the structural features of cascades includes studying their structure and quantifying cascade networks' properties. According to Liben-Nowell and Kleinberg [43], a better understanding of the properties of the structure of cascades leads to better dissemination models. Table 1 lists the structural features of cascades that are categorised into three categories. The first category are cascade-centric features; these features are computed on the cascade level as a whole. The significance of each of these features is as follow:

1. **Depth, range, distance to the root:** Indicates the shape of a cascade, and how far it travels away from the source within the network. When all distances to the root are gathered, they can help assessing whether a cascade is shallow or deep [43, 24, 38, 9, 55, 3, 25, 21, 17, 6].
2. **Width**: Indicates the extent to which a cascade is narrow or wide. It gives hints about the factors that make a message quite popular at one stage within the cascade [43].
3. **The fraction of nodes with exactly one child**: Indicates missing or unsuccessful cascade event [43, 6].
4. **Scale**: Indicates how popular/interesting a message gets soon after its first appearance [55, 6].
5. **Wiener index**: Gives an indication of the cascade shape, the higher the Wiener index, the more viral the cascade. Cascades with low Wiener index resemble a star shape, where there are few hubs that create the cascade. The Wiener index increases with the increase in cascade size [18, 7, 5].
6. **The percentage of adoption per depth**: Counting the percentage of adoptions within one degree of a root could indicate whether epidemic-like cascade occurs in the dataset, i.e. if the majority of adoptions recorded in the dataset are within the first few degrees from a root, then one could conclude that most cascades are shallow and small [25, 7, 6].
7. **Number of nodes at depth = 1**: Nodes (users) at depth 1 are the ones who share directly from the author, meaning that they were exposed to the authors post directly. It might be that they arrive via external resources or direct links. Although there is a possibility that users click on the original post and share from the author rather than from user they receive the post from [21, 6].
8. **Connectivity Rate**: Shows whether an edge exists between any two nodes in the cascade. It is useful to examine whether users get their information from the social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree [52].
9. **Root Fragment Rate**: Shows whether each node in the cascade is actually linked to the root or not. It is useful to examine whether users get their information from social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree [52].
10. **Diameter**: Shows whether cascades are deep or shallow [42, 52].

The second category is node-centric structural features, which are computed on nodes level. There are two features in this category: the branching factors

Table 2: Temporal Features

Feature	**Definition**
Time passed since message published [21]	How many times a particular message has been passed in relation to the time since it was published.
Speed [55]	Detecting whether and when the first cascade will occur (depth = 1).
Time lag between posting and first reshare, elapsed time [38, 17, 6]	The difference between posting time and the first reshare.
Time lag between two sharing events [38]	The difference between two nodes in a cascade.
The number of spikes/peaks [31, 19]	Spikes refer to high-volume of cascading activities that occur in a short period during the lifetime of a cascade.
Cascading density throughout lifetime [31, 6]	The timeline of a cascade, it shows the number of cascading activities per day.
Maximum time between reshares [19, 6]	The maximum time difference between reshares.
Cascade growth/cascade popularity [41, 3, 21, 7, 6]	The relation between the growth in cascade size through time. The rate at which cascades gain their size (i.e. popularity).
Recurrence [19]	Recurrence occurs if a cascade has at least two peaks in addition to other conditions.

and the subcascade size and they both measure individual's influence on the overall cascade [31, 24, 21, 5, 6]. However, there is a difference between the two, as the branching factor estimates the immediate influence, the subcascade size estimates the overall influence of one individual on the cascade. It is important to take the two measures into account as some nodes might have a small branching factor but their subcascade might be very large [6].

The last structural feature is the frequency of distinct cascade structures. It helps to detect if there is a repeated cascade pattern, which can be investigated later. When combined with depth, it could help draw some conclusions about the shape of the cascade and how far it branches [42, 41, 25, 17].

4.2 Temporal Features

There are two approaches to analyse the temporal aspect of cascades. The first tracks and describes existing cascades' temporal features, e.g. how fast information spreads, for how long trendy content keeps its popularity, and the overall growth of cascades over time, such as: whether cascades show patterns like 'burstiness' or sparks. The other line of research uses cascade's temporal patterns to either predict or model the cascade's future popularity. Most of these studies do use the word 'cascade', because they are concerned with the temporal

aspect of the diffusion of online content. However, the underlying structure of online content diffusion is an implicit cascade network.

Table 2 lists a number of cascades' temporal features, their significance is as follow:

1. **Time passed since message published**: Shows the growth of cascade and the fade of interest in the message over time [21].
2. **Speed**: Indicates how fast users would be influenced to spread the message or generally react using other means of interaction like reply or mention [55].
3. **Time lag between posting and first reshare, elapsed time**: Measures the resharability of content: the larger the lag the less likely a content will be reshared [38, 17, 6].
4. **Time lag between two sharing events**: Shows the speed at which a cascade occurs in relation to the distance between nodes, i.e. sharing events [38].
5. **The number of spikes/peaks**: Measures the degree to which a cascade provokes high volume of cascading during its lifetime [31, 19].
6. **Cascading density throughout lifetime**: Helps assessing the temporal patterns of diffusion, whether it has spikes or maintains a steady growth. [31, 6]
7. **Maximum time between reshares**: Indicates the maximum idleness period within a cascade [19, 6]
8. **Cascade growth/cascade popularity**: Helps to show whether a cascade size grows linearly as time passes or in different ways. This helps detect whether the growth in cascade size occurs in short intervals or whether it grows with time. It also shows the periods of idleness and spikes in the cascade timeline [41, 3, 21, 7, 6].
9. **Recurrence**: Helps identifying cascades that regain their popularity after a period of idleness [19].

5 Conclusions

In this paper, we presented a survey of cascades, cascade networks and cascade features. Our aim was to investigate these subjects while considering two aspects: the content type and the platform. The main message this paper conveys is that content type is significant in the process of constructing and analysing cascades. Not only that, but, content type has an impact on the approaches used to collect the datasets as well. In addition, the survey of cascade features will be useful for researchers who would like to study cascades, as it will give them an overall overview of the measures that have been used in the literature including their significance as cascades estimators. These features can be used for several purposes related to cascades: quantifying, modelling and predicting their future growth.

Bibliography

[1] World Wide Web Timeline (2014), http://www.pewinternet.org/2014/03/11/world-wide-web-timeline

[2] Adamic, L., Lento, T., Adar, E., Ng, P.: Information Evolution in Social Networks. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WDSM'16). pp. 473–482. ACM, New York, NY, USA (2016)

[3] Adamic, L.A., Lento, T.M., Fiore, A.T.: How You Met Me. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM) (2012)

[4] Adar, E., Adamic, L.: Tracking Information Epidemics in Blogspace. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI '05). pp. 207–214. IEEE Computer Society (2005)

[5] Alrajebah, N.: Investigating the Structural Characteristics of Cascades on Tumblr. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM 2015). pp. 910–917. ACM (2015)

[6] Alrajebah, N., Carr, L., Luczak-roesch, M., Tiropanis, T.: Deconstructing Diffusion on Tumblr : Structural and Temporal Aspects. In: Proceedings of the 9th ACM Conference on Web Science. ACM (in press)

[7] Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., Tiwari, M.: Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily. In: Proceedings of the 24th International Conference on World Wide Web (WWW'15). pp. 66–76. ACM (2015)

[8] Antoniades, D., Dovrolis, C.: Co-evolutionary dynamics in social networks: A case study of Twitter. Computational Social Networks 2(14) (2015)

[9] Bakshy, E., Hofman, J., Mason, W.A., Watts, D.J.: Everyone's an Influencer: Quantifying Influence on Twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). pp. 65–74. ACM (2011)

[10] Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The Role of Social Networks in Information Diffusion. In: Proceedings of the 21st international conference on World Wide Web (WWW '12). pp. 519–528. ACM, Lyon, France (2012)

[11] Berners-Lee, T.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. HarperInformation (2000)

[12] Bhattacharya, D., Ram, S.: Sharing news articles using 140 characters: A diffusion analysis on twitter. In: Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM '12). pp. 966–971. IEEE Computer Society (2012)

[13] Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy 5, 992–1026 (1992)

[14] Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., Wallach, D.S.: Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. ACM Transactions on Internet Technology 15(1), 24 (2015)
[15] danah Boyd, Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: 2010 43rd Hawaii International Conference on System Sciences (HICSS). pp. 1–10. IEEE Computer Society (2010)
[16] Çelen, B., Kariv, S.: Distinguishing informational cascades from herd behavior in the laboratory. The American Economic Review 94(3), 484–498 (2004)
[17] Chang, Y., Tang, L., Inagaki, Y., Liu, Y.: What is Tumblr: A Statistical Overview and Comparison. SIGKDD Explorations 16(1), 21–29 (2014)
[18] Cheng, J., Adamic, L.A., Dow, P.A., Kleinberg, J., Leskovec, J.: Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web (WWW'14). pp. 925–935. ACM, Seoul, Korea (2014)
[19] Cheng, J., Adamic, L.A., Kleinberg, J., Leskovec, J.: Do Cascades Recur? In: Proceedings of the 25th International Conference on World Wide Web (WWW'16). pp. 671–681. ACM (2016)
[20] Cheong, M., Lee, V.: Twittering for earth: A study on the impact of microblogging activism on earth hour 2009 in Australia. In: Intelligent Information and Database Systems (ACIIDS 2010). vol. 5991 LNAI, pp. 114–123. Springer Berlin Heidelberg (2010)
[21] Dow, P., Adamic, L., Friggeri, A.: The Anatomy of Large Facebook Cascades. In: Proceedings of the Seventh International Conference on Weblogs and Social Media, (ICWSM). pp. 145–154. AAAI, Cambridge, Massachusetts, USA (2013)
[22] Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)
[23] Farajtabar, M., Gomez-Rodriguez, M., Wang, Y., Li, S., Zha, H., Song, L.: Co-evolutionary Dynamics of Information Diffusion and Network Structure. In: Proceedings of the 24th International Conference on World Wide Web (WWW'15). pp. 619–620. ACM (2015)
[24] Galuba, W., Aberer, K.: Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In: Proceedings of the 3rd Wonference on Online Social Networks (WOSN'10). pp. 1–9. USENIX Association, Boston, MA (2010)
[25] Goel, S., Watts, D., Goldstein, D.: The structure of online diffusion networks. In: Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012). vol. 1, pp. 623–638. ACM (2012)
[26] Goldenberg, J., Libai, B., Muller, E.: Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters 12(3), 211–223 (2001)
[27] Gomez Rodriguez, M., Leskovec, J., Krause, A.: Inferring Networks of Diffusion and Influence. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - (KDD '10). pp. 1019–1028. ACM (2010)

[28] Gomez Rodriguez, M., Leskovec, J., Schölkopf, B.: Structure and dynamics of information pathways in online media. In: Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13). p. 23. ACM (2013)
[29] Grabner-Kräuter, S.: Web 2.0 Social Networks: The Role of Trust. Journal of Business Ethics 90(SUPPL. 4), 505–522 (2009)
[30] Granovetter, M.S.: Threshold Models of Collective Behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
[31] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion Through Blogspace. In: Proceedings of the 13th international conference on World Wide Web (WWW '04). pp. 491–501. ACM (2004)
[32] Guille, A., Hacid, H., Favre, C., Zighed, D.: Information Diffusion in Online Social Networks: A Survey. SIGMOD Record 42(2), 17–28 (2013)
[33] Heidemann, J., Klier, M., Probst, F.: Online social networks: A survey of a global phenomenon. Computer Networks 56(18), 3866–3878 (2012)
[34] Herring, S.C.S.: Computer-mediated communication: linguistic, social, and cross-cultural perspectives, vol. 39. John Benjamins Publishing (1996)
[35] Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management 6, 248 (2009)
[36] Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 53(1), 59–68 (2010)
[37] Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the Spread of Influence through a Social Network. In: Proceedings of the ninth ACM international conference on Knowledge discovery and data mining (KDD'03). pp. 137–146. ACM (2003)
[38] Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web (WWW '10). pp. 591–600. ACM (2010)
[39] Lai, L.S.L., Turban, E.: Groups formation and operations in the web 2.0 environment and social networks. Group Decision and Negotiation 17(5), 387–402 (2008)
[40] Lerman, K., Ghosh, R.: Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In: Fourth International AAAI Conference on Weblogs and Social Media. pp. 90–97. AAAI (2010)
[41] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Patterns of cascading behavior in large blog graphs. In: Proceedings of the 2007 SIAM international conference on data mining. pp. 551–556. SIAM (2007)
[42] Leskovec, J., Singh, A., Kleinberg, J.: Patterns of influence in a recommendation network. In: Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'06). pp. 380–389. Springer-Verlag, Singapore (2006)
[43] Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using Internet chain-letter data. Proceedings of the National Academy of Sciences 105(12), 4633–4638 (2008)

[44] Ma, Z., Sun, A., Cong, G.: On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology 64(7), 1399–1410 (2013)
[45] McBride, K.: Journalism and public shaming: Some guidelines (2015), http://www.poynter.org/2015/journalism-and-public-shaming-some-guidelines/326097/
[46] Myers, S., Leskovec, J.: The Bursty Dynamics of the Twitter Information Network. In: Proceedings of the 23rd International Conference on World Wide Web (WWW '14). pp. 913–923. ACM (2014)
[47] Myers, S.a., Leskovec, J.: Clash of the Contagions: Cooperation and Competition in Information Diffusion. In: Proceeding of IEEE 12th International Conference on Data Mining. pp. 539–548. IEEE (2012)
[48] Newman, M.E.J.: Networks: an introduction. Oxford University Press (2010)
[49] O'Reilly, T.: What Is Web 2.0 (2005), http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html
[50] Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! Predicting Message Propagation in Twitter. In: Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM). pp. 586–589. AAAI (2011)
[51] Scott, J.: Network analysis. In: Darity, W.A. (ed.) International Encyclopaedia of the Social Sciences. Macmillan Reference USA (2008)
[52] Taxidou, I., Fischer, P.M.: Online Analysis of Information Diffusion in Twitter. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW Companion '14). pp. 1313–1318. ACM (2014)
[53] Weng, L., Flammini, A., Vespignani, A., Menczer, F.: Competition among memes in a world with limited attention. Scientific Reports 2, 1–9 (2012)
[54] Xu, J., Compton, R., Lu, T.C., Allen, D.: Rolling through Tumblr : Characterizing Behavioral Patterns of the Microblogging Platform. In: Proceedings of the 2014 ACM Conference on Web Science (WebSci '14). pp. 13–22. ACM (2014)
[55] Yang, J., Counts, S.: Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. In: Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM '10). pp. 355–358. AAAI (2010)
[56] Yang, L., Sun, T., Zhang, M., Mei, Q.: We know what@ you# tag: does the dual role affect hashtag adoption? In: Proceedings of the 21st international conference on World Wide Web (WWW '12). pp. 261–270. ACM (2012)