Aragón *et al. Journal of Internet Services and Applications* (2017) 8:15
DOI 10.1186/s13174-017-0066-z

Journal of Internet Services and Applications

RESEARCH Open Access

Generative models of online discussion threads: state of the art and research challenges

Pablo Aragón[1,2], Vicenç Gómez[1], David García[3] and Andreas Kaltenbrunner[1,4*]

Abstract

Online discussion in form of written comments is a core component of many social media platforms. It has attracted increasing attention from academia, mainly because theories from social sciences can be explored at an unprecedented scale. This interest has led to the development of statistical models which are able to characterize the dynamics of threaded online conversations.

In this paper, we review research on statistical modeling of online discussions, in particular, we describe current generative models of the structure and growth of discussion threads. These are parametrized network formation models that are able to generate synthetic discussion threads that reproduce certain features of the real discussions present in different online platforms. We aim to provide a clear overview of the state of the art and to motivate future work in this relevant research field.

Keywords: Online discussion, Computer-mediated communication, Discussion threads, Computational social science, Social media

1 Introduction

The success of the Internet has generated a wide variety of platforms for computer-mediated communication. Through these, online discussion has become an important part of the communication of modern societies and the interest in this form of online discussion is still growing at the moment this manuscript is written [1]. Discussions on the Internet commonly occur as a exchange of written messages among two or more participants. In this way, conversations are often represented as threads, which are initiated by a user posting a starting message (hereafter post) and then users send replies to either the post or the existing replies. Therefore, given this sequential posting behavior, online discussion threads follow a tree network structure. Previous and recent research has analyzed this network structure of online discussions for different and relevant purposes, e.g. the resolution of problems in e-learning platforms [2], the response of online communities to natural disasters [3], the spread of rumors in social media [4], etc.

In point of fact, we are witnessing the age of computational social science [5] where the collection and analysis of online data (e.g. data from online discussion threads) provide interesting insights on human behavior. Although some voices claimed that data-driven approaches will make the scientific method obsolete [6], statistical and theoretical motivated models are needed to determine which are the social factors explaining these network structures. In the context of online discussions, different modeling approaches have been proposed to identify the governing mechanisms of the structure of threads. Statistical models of this type are aimed to reproduce the growth of discussion threads through different features, often related to human behavior. This is why these are usually called *generative* models: they do not only estimate the statistical significance of their corresponding features but also reproduce the temporal arrival patterns of messages that form a discussion thread.

Despite the recent effort on surveying statistical graph models for social networks [7, 8], to our best knowledge, there is no formal review of generative models of

*Correspondence: kaltenbrunner@gmail.com
[1]Universitat Pompeu Fabra, Barcelona, Spain
[4]NTENT, Av. Diagonal 220, 08018 Barcelona, Spain
Full list of author information is available at the end of the article

online discussion threads. The aforementioned relevance of online discussions and the informative value of statistical models to explain human behavior motivate us to fill this gap in the literature by presenting the following survey. Thus, the purpose of this survey is threefold:

- to provide the reader with a structured comparison of the several existing generative modeling efforts for online discussions threads,
- to provide guidelines on how to evaluate and extend these models,
- and to sketch their existing and potential future applications.

These applications span areas as specific as evaluating the impact of even minor platform changes over comparing different user communities towards generic studies of human online communication behavior .

Before we start introducing the different modeling approaches, we present in Section 2 a historical overview of the different paradigms and platforms for online discussion and describe in Section 3 social theories which are present in threaded online discussions but are addressed only in part by the current generative models. We introduce then the best practices and guidelines when modeling the structure of these online discussion threads in Section 4 to continue in Section 5 with the statistical models which have been proposed in this respect. In Section 6 we discuss the practical applications of these generative models. Finally, we identify in Section 7 some open research challenges and conclude in Section 8.

2 A brief historical overview of online discussion platforms

The history of online discussion helps to understand the history of the Internet itself. Thus, the way online discussions are structured can be explained through the different paradigms of Internet conversation media. Previous work on evolutionary perspective of Internet conversation media has indicated email services as the first platforms in which users were able to exchange messages across computer networks [9]. This type of communication dates back to the 1960s and is still one of the most common channels for online discussion. Although the first email systems implemented *synchronous* communication only, requiring to both sender and receiver users be online at the same time, they moved rapidly towards the current store-and-forward model of *asynchronous* communication. These two paradigms are the basis of the taxonomy of Internet conversation media presented in Fig. 1:

- *Synchronous communication.* Platforms of this type could be characterized by the language form:
  - text: instant messengers and chats,
  - audio: voice call and audioconference,
  - video: video call and videoconference.

  The difference between the two types in each form is the privacy of the discussion, i.e., private (between two users) or group (between more than two users). For text-based discussions, chat systems were developed to allow users to exchange of messages (usually short) in real-time. Despite the first chat system was developed in the 1970s [10], online chats became more and more popular only after the Internet Relay Chat (IRC) protocol was created in 1988 [11].
- *Asynchronous communication.* The next platforms to be developed could be characterized by the level of privacy. For private conversations, the Short Message Service (SMS) protocols were defined in the '80s [12]. Discussion in these media are structured as an exchange of messages between pairs of users which might be represented as a chain. For group conversations (e.g., groups messages, discussion lists), bulletin board systems (BBS) were the next step. In BBSs, users log in to a computer system mainly for reading news/bulletins (newsgroups) and exchanging messages with other users through public message boards. This type of communication is the origin of Usenet, a worldwide distributed asynchronous discussion system [13] which introduced the notion of discussion threads to organize newsgroup conversation.

The irruption of the World Wide Web [14] revolutionized the existing online platforms for both synchronous and asynchronous communication. On the one hand, many web services and browsers built in IRC clients. On the other hand, asynchronous discussion platforms evolved from BBSs and Usenet networks into web-based forums dedicated to very different types of themes, e.g. politics, technology, etc. In online forums, conversations follow the thread structure inherited from Usenet. We should note that, in comparison to earlier asynchronous communication systems, online forums still have immense popularity on the Internet (e.g., 4chan forum has over 22 million monthly visitors worldwide[1]).

In the entry into the 21st century, the Web evolved by the outbreak of blogs and social media: "weblogs turned from an ease-of-publishing phenomenon into a conversational mess of overlapping communities" [15]. Online discussion in the blogosphere, in which messages are typically structured as replies to the post, was accompanied by the emergence of online social networking sites which built new spaces of online discussion for different types of themes and purposes. One of the most popular platforms, also reflected by the interest from academia, is Facebook. Users in this social networking site are able to discuss

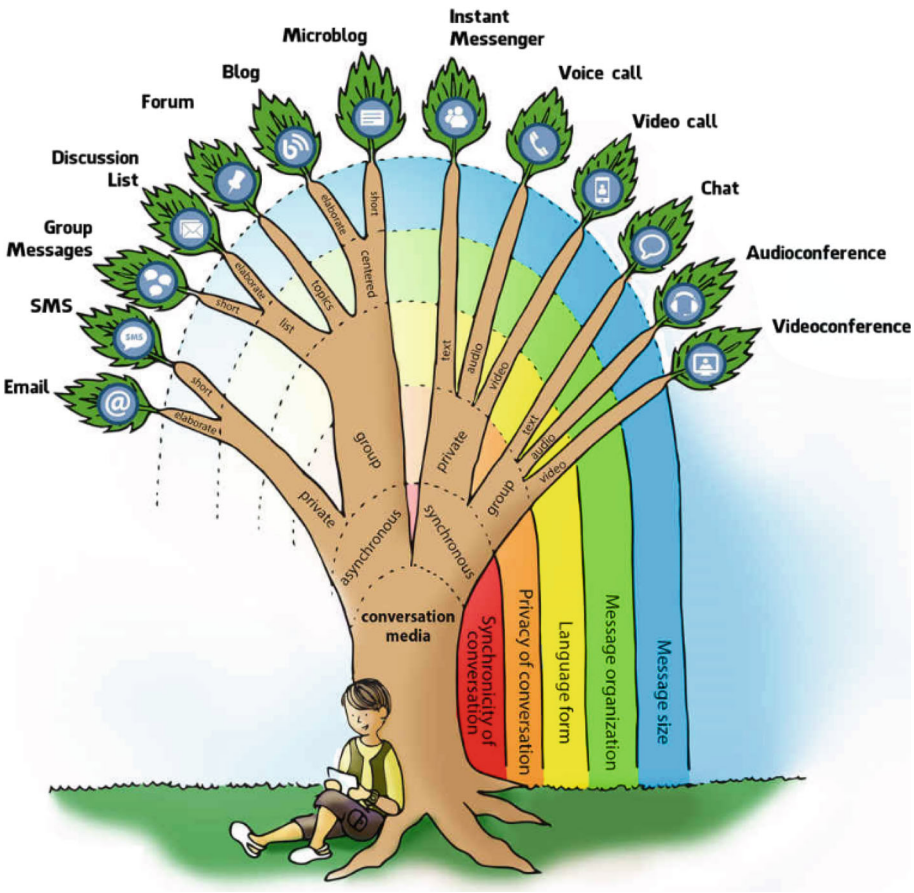

Fig. 1 Taxonomy of Internet conversation media (Source: [9]). Our article surveys generative models of online discussion threads, i.e. conversations from platforms corresponding to asynchronous communication (leafs of the left main branch of the tree)

with other users, mostly with their network of friends. We should note that this platform was preceded by similar sites in which users also maintained conversations with their personal network of online friends, e.g. MySpace, Friendster, and Orkut. Another type of social media platform which has received great attention in research are the microblogging services. Twitter is the most studied site of this type, arguably due to the massive usage of this platform in recent years and also because most of the data about online discussion is publicly available and easily accessible. Finally, we should note that many other social platforms which were developed for social news (e.g. Digg, Reddit), multimedia content (e.g. Flickr and Instagram for photos, Youtube and Vimeo for videos), peer production (e.g. Wikipedia, Github), online education (e.g. Coursera) have incorporated discussions as a essential component of the platforms themselves.

As we specified in the introduction, this is a survey on generative models of online discussion threads. Therefore, the rest of the paper only focuses on text-based synchronous and asynchronous Internet conversation media, i.e., instant messengers, chats, and the leafs of the left main branch of the tree in Fig. 1.

3 Social theories in online discussions

Among the many diverse topics explored in the literature, an important fraction of research in online discussion has explored well-known theories from sociology and social psychology. In this section we review how three relevant social theories explain behavior in online discussions: homophily, social influence, and emotional contagion. From these three theories, social influence has already found its way into the generative models presented in Section 5, the other two are presented here as a reminder of open research challenges to be addressed in Section 7.

3.1 Homophily

"Birds of a feather flock together" is a proverb that captures the principle of *homophily*: the contact between similar people is more likely than among dissimilar ones [16]. Sociologists have systematically studied homophily since the middle of the 20th century, starting from the analysis

of friendships across housing communities [17]. Among the causes of homophily are the foci of formation, for example segregated neighborhoods; the isomorphism of network positions, for example when company directors become friends due to their similar organizational roles; and selective tie dissolution, such as professional relationships surviving longer than leisure ones [16]. A substantial body of evidence exists for the presence of homophily in offline social networks, where people tend to form links of all sorts, such as friendship, support, or mere contact links, with other people of similar age, ethnicity, gender, religion, and social class.

Homophily in social networks is commonly tested through the topological principle of assortativity, i.e. that the attributes at the ends of social links are correlated [18]. While homophily implies assortativity, assortativity can also be a manifestation of social contagion or peer influence when attributes are acquired (such as religion or occupation) rather than ascribed (like race or age). In the context of online discussions, the creation of contacts between users happens at a much faster timescale than the changes of individual attributes, such as age or political views. This difference between timescales makes homophily a frequent plausible explanation for the origin of assortativity in online discussions. On the contrary, other phenomena at similar timescales as the online discussion, like belonging to an online group or sharing content, require additional considerations to disentangle homophily from social contagion [19].

Homophily with respect to demographic characteristics has been observed in a wide variety of online platforms and media. Online discussion in MSN Messenger exhibits homophily with respect to interests, age, and location [20]. In MySpace, commenting across profiles shows homophily with respect to a wide variety of demographic factors [21], including ethnicity, religion, age, and marital status. Sexual orientation shows a pattern of online homophily as well: homosexual men are more likely to be friends with homosexual men in Facebook [22], and in the now disappeared Friendster social network [23], lending the ground for social inference of sexual orientation based on digital traces. It is worth noting that, with respect to gender, online interaction exhibits a pattern of heterophily, i.e. users tend to interact with users of the opposite gender. This pattern is present in various social networks [20, 21, 24] and suggests the role of this kind of communication for mating and romantic purposes.

Homophily is also present in online interaction with respect to psychological behavior and traits. Personality traits, in particular emotional stability, openness, and extraversion, show homophily patterns in Facebook friendship links [25]. Another noticeable aspect of homophily is happiness, which is known to be assortative in offline social networks [26]. Using sentiment analysis methods, subjective well-being and happiness has been found in Twitter when analyzing bidirectional reply links [27] and bidirectional follower links [28].

Homophily with respect to political orientation is often analyzed in relation to the more general phenomenon of polarization. Links between US political blogs in 2004 have shown a strong pattern of homophily along the liberal-conservative spectrum [29], a phenomenon that is often discussed in relation to the existence of echo chambers [30], the filter bubble [31], and selective exposure to political information [32]. Online interaction in Twitter shows clear signs of homophily with respect to political orientation, which can be used to infer user political alignment by analyzing the follower network [33]. Retweet networks show homophily with respect to user party alignments in the US [34], in Germany [35] and in Spain [36]. The user mention networks show weaker homophily in all three cases. Beyond Twitter, recent work on online political networks [37] show that network layers with positive connotation (supports and likes) display stronger patterns of homophily with respect to party alignment than the layer of comments, where no homophily is present. This finding is consistent with studies on Wikipedia [38] in which editors who display their party alignment on their profile show homophily with respect to this alignment when interacting through their user walls, but not through common discussions in talk pages (see Fig. 2). Homophily in Wikipedia has been also found with respect to user experience, i.e. very active preferred to interact with inexperienced users [39, 40].

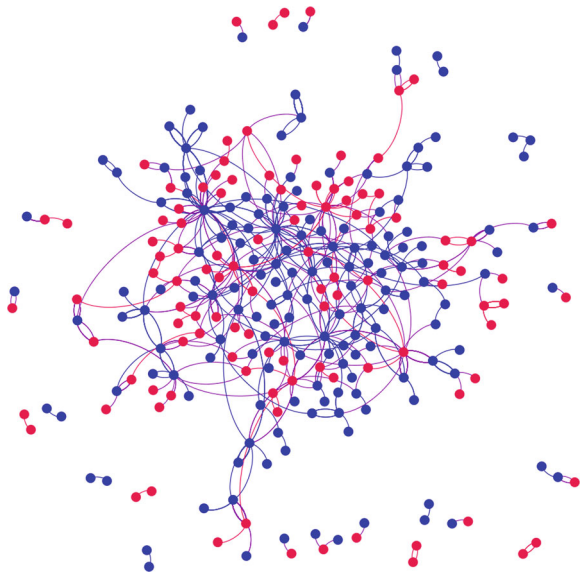

Fig. 2 The reply network in Wikipedia talk pages between Democrats (blue nodes) and Republicans (red nodes) shows no homophily with respect to party alignment. Data from [38]

3.2 Social influence

The behavioral change of an individual caused by the interaction with other individuals is called social influence. The reasons and forms of these attitude changes are a well-studied topic in social psychology. Social influence can be either informational (people need to be right) or normative (people need to be liked) [41]. According to [42], the way in which people influenced one another are categorizable in three different processes: people publicly agree with others while privately dissent (compliance), people are influenced by the people with whom they feel identified (identification), and people agree both publicly and privately without coercion from others (internalization). Indeed, popular models of collective behavior are precisely based on the premise that people are affected by the influence of others [43].

Social influence is considered a typical theory of social networks [44–46]. Because the network structures produced by social influence are similar to the ones generated by homophily (e.g. temporal clustering), some studies have developed frameworks to distinguish both theories in diffusion networks [47]. Indeed, a large fraction of research has focused on social influence for message propagation in online platforms, e.g. [48–50]. However, some other studies have explicitly evaluated the role of social influence in online discussion.

One of the more comprehensive analyses of social influence in online discussion [51] distinguished seven types of platforms categorized as network-based communities (email lists, bulletin board systems and Usenet newsgroups) and small-group-based communities (online-chat systems, web-based chat rooms, multiplayer virtual games and multiuser domains). Results show that the type of community influenced how to convey member information to other users. In particular, because many users in network-based communities are strangers at first, reputation seemed to play a key role for fostering social interactions and, therefore, might explain why many of these platforms (e.g. Slashdot) display the contribution history of users.

An analysis of the exchange of email messages between thirty-two students in an online learning environment has shown that social influence affected the discussions, in particular, users were more inclined to follow social recommendations made by highly central users than those by peripheral ones [52]. In contrast, a study of different blogging platforms found that community identification might be the motivation of users to participate in blogs, while the influence of social norm was not relevant [53]. A later study of the messages by thousands of participants across 16 Google Groups concluded that activity and tenure of discussion within a group were related to the ability to influence others [54].

In microblogging services, a study of the online discussions on Twitter about the Haiti earthquake revealed that when the percentage of one's friends joining the discussion increases, the likelihood that the user also participates increases too [55]. Social influence has been also detected on Youtube in a experiment in which the comments of videos were proven to affect the evaluation of the videos' owners [56].

3.3 Emotional contagion

Emotional contagion refers to the process by which individual emotions are triggered by similar emotional states in other individuals [57]. The study of emotional contagion poses emotions as physiological states that make their subject infectious while interacting with others. The most traditional perspective to this phenomenon builds on principles of mimicry, by which the facial feedback of emotions can transmit emotional states without an individual noticing [58]. More recent theoretical approaches extend the modalities and scope of social aspects of emotions. The hyperlens model of emotions [59] defines social regulation of emotions as a generalized case of emotional contagion, which can happen unconsciously, like in the case of mimicry, or consciously, such as in a discussion or sharing of emotional experiences. Online interaction plays an important role with respect to emotional contagion: extensive research has shown that emotional contagion is present in computer-mediated communication as much as in face to face communication [60]. Furthermore, recent research has experimentally quantified the dynamics of emotions while reading and writing in forum threads [61], illustrating that emotional contagion in online discussions is a very present phenomenon.

The digital traces left in online interaction allow to analyze emotional contagion at much larger scales than in laboratory studies. A large amount of Facebook data was processed with psycholinguistic methods to detect emotional contagion, by using weather as an instrumental variable [62]. Furthermore, manipulations of the selection of content seen by Facebook users led to their emotions moving in the predicted direction [63]. While these two studies show that digital trace data is powerful to measure emotional contagion, they suffer two important limitations: first, the data generated during experiments in a private company like Facebook is not available for public research inspection, and second, clear ethical concerns rise from the manipulation of emotions under weak consent scenarios, such as the user terms of Facebook. Observational research on Twitter has also shown the existence of emotional contagion without those limitations, illustrating through sentiment analysis the correlations between emotional expression at the endpoints of online interaction [64]. In addition, the analysis of emotional expression in the Chinese website Weibo, very

similar to Twitter, shows asymmetric properties of emotional contagion: Anger seems to be more contagious than joy [65]. In the case of Wikipedia, a study of messages from article talk pages found that editors tended to interact with editors with a similar emotional style [66] (see Fig. 3). Although this might be an effect of emotional contagion, the authors of the study suggested that this pattern might be the result of emotional and linguistic homophily.

Emotional contagion in online discussions has a wide range of consequences for online and offline life. Emotions expressed in forums lead to the emotional activation of users reading them, increasing the chance of participating in online discussions [61]. This way, emotional expression prolongs online discussions, a phenomenon which has been observed for the case of negative emotions in BBC forums [67]. Similarly, retweets are more likely to express strong and bipolar emotions than normal tweets [68], indicating that emotions can play a role in information spreading. One example is the signature of spreading of negative emotions, which is generally wider than positive emotions [69]. A second example is the pattern of cascades of online activity related to political movements, where emotionally charged discussions spread larger and further in the social network [70].

Finally, collective emotions are emotional states simultaneously shared by large amounts of individuals, often as a result of emotional contagion [71]. Various online forums and discussion media show signs of collective emotions, with clusters of similarly emotional posts [72].

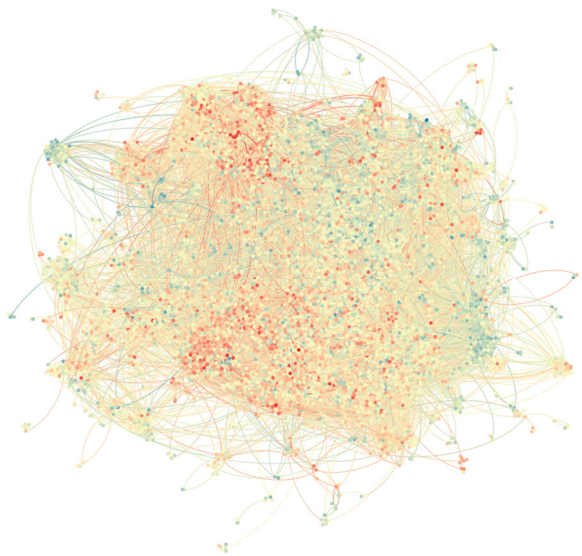

Fig. 3 Reply network of users on Wikipedia article talk pages. The color of nodes expresses the proportion of words expressing anger (from blue to red). Assortativity observed in this network (e.g. clusters of red nodes) might be explained by either homophily or emotional contagion. Data from [66]

Collective emotional states show persistence patterns that are built on the emotional interaction between individuals. Thus, the analysis of emotions in real time chats shows this effect, as emotional persistence builds up on the memory and social interaction between participants of a group chat [73].

4 Data-driven modeling of online discussion threads

In this review, we refer to modeling as defining a mathematical description of a process that generates online discussion threads, as illustrated in Fig. 4. We consider data-driven models, that is, models that try to capture some phenomena of interest of a given dataset [74, 75]. Such models are constrained by the nature of the data and the type of phenomena that they try to explain. For example, a model can be defined at the fine-grained level of the individual text of a comment or it may abstract an entire conversation from the content of the messages. Also, it can describe the precise timing when comments are send/received or it can completely disregard any temporal aspect. In this review, we focus on models that incorporate a fundamental ingredient of online discussion threads: their reply structure.

Contrary to purely descriptive approaches, generative models are also able to produce instances of the objects of interest, in our case, synthetic instances of discussion threads. Generative models provide more insights and explain better the formation process of online discussion threads than purely descriptive approaches [76].

The behavior of a model depends on its parameters θ, that are adjusted to fit the data. It is important to differentiate between fully data-driven models and models that are largely constrained using prior knowledge. Fully data-driven models usually depend on a large number of parameters and are used as *black-box* models. They are typically trained end-to-end to optimize some measure of predictive performance. Conversely, parsimonious models try to explain phenomena with as few parameters as possible. An example would be a model with a single parameter that is able to generate conversation threads with the same degree distribution than the real ones. Although the latter type of models may perform worse than fully data-driven models in terms of predicting power, they tend to be more interpretable [77] and can thus provide a better understanding of the governing mechanisms of online discussions.

To estimate the model parameters, the most common approach is to optimize a likelihood function, which quantifies how good the model explains the data as a function of its parameters [75]. While this optimization has analytical solution for very simple models, it can be in general computationally challenging. The complexity of such an optimization depends on the model complexity. Models

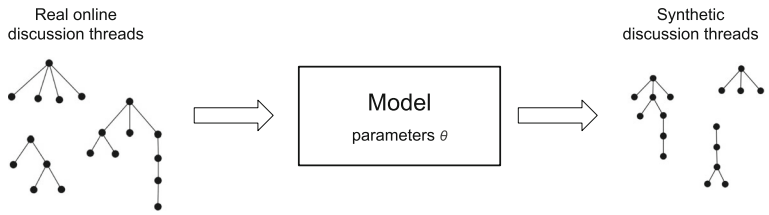

Fig. 4 Modeling approach considered in this review: the model (box in the middle) represents a mechanism or procedure that describes how discussion threads are formed. It is usually governed by a set of parameters θ which are typically learned from real data composed of real discussion threads. This learning step involves some type of optimization. For given parameters θ, the model can be used to generate synthetic threads that reproduce the properties of the real discussion threads

with a large number of parameters compared to the size of the data may fail to generalize and their predictions may be not valid for new data. In these cases, adding some form of regularization and partitioning the dataset into different subsets for training, validation and testing (cross-validation) helps. In any case, it is necessary to take into account the statistical assumptions in the data generating process. For example, whether data points are independently distributed and under stationary conditions between training and testing conditions, which is often not the case.

A model is expected to be *identifiable*, i.e. as the number of data increases, the true parameter values must converge. For a learned identifiable model, distinct parameter values θ should correspond to distinct models. In contrast, a model is said to be *non-identifiable* when different parameter values result in the model. This can occur, for example, when flat directions exist in the likelihood landscape. To evaluate identifiability, as a sanity check, a good strategy is to:

1. generate synthetic data with some parameter values θ^* from the estimation,
2. train the model with those data,
3. evaluate whether the model estimates consistently the same parameter values θ^*.

This could be done for different choices of θ^*.

The validation of generative models, in particular, network formation models, typically examines whether the structural properties (e.g. size, depth, width) of generated data are comparable to the original properties of the empirical data. Usually, statistical tests are used for this purpose, e.g. the Kolmogorov-Smirnov (KS) test [78–80], which measures the maximum punctual distance between the empirical cumulative distribution function (CDF) $F_e(x)$ and the CDF of the generated synthetic data $F_g(x)$, defined both on observations x of the structural property of interest:

$$\text{KS-stat} = \sup_x |F_e(x) - F_g(x)|.$$

The use of such tests provides strong statistical evidence favoring a model. Very often, however, these tests can be too strict due to finite-size effects or other artifacts present in the data. In these cases, an alternative qualitative validation, for example, using visualization techniques, can be also satisfactory.

An example of the types of structures that a generative model for discussion threads can generate is shown in Fig. 5. In this hierarchical structure, or reply tree, the nodes are messages (the post is the root), and the edges are the directed reply relationships from a reply to the message it replies to (known as parent). At the node level, the number of children (also referred to degree or branching) is the number of replies the message received, and its distance to the root (in terms of number of edges) is the reply level. Some *structural properties* of discussion threads are:

- *size*: the number of messages,
- *width*: the maximum number of messages at any reply level,
- *depth*: the length of the largest exchange of messages,
- *users*: if the message authorship is known, number of users who authored at least one message.

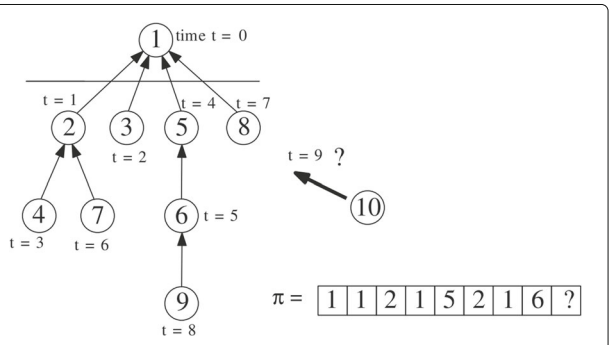

Fig. 5 Example of a discussion thread represented as a tree: at time-step $t = 9$, node (comment) number 10 is added to the thread. The parent of each node is sequentially stored in vector π. Source: [85]

4.1 Typical features of online discussion threads

To better identify features of online discussion, we introduce below the most common ones, using the previous illustrative discussion thread. *Popularity* is a classic feature which expresses that the more connected a node is (the most commented messages), the more likely a node is to attract new edges (new replies). This feature is usually introduced trough the preferential attachment process [81], also referred to as the Yule process [82]. This process is a common property of many social networks and establishes that the probability that one of the links of a new node connects to certain node depends on its degree. Thus, node 1 in Fig. 5 is the most popular message, followed by node 2, then nodes 5 and 6 and, finally, the rest of the nodes. Besides the popularity of nodes, it is usually expected that the newest comments are the most attractive messages. Therefore, many models introduce *novelty* as a feature, i.e. nodes 1 and 2 are the most popular ones but also the oldest ones, which might reduce the arrival of new replies. Moreover, notice an important difference: node 1 is the initial post while node 2 is a comment, like the rest of the nodes. Some users might be interested in replying directly to the post while some other users might be interested in replying to comments and getting engaged in the discussion. This is the reason why certain models establish a *root-bias* as a feature. Some models also consider that the result of users getting engaged in discussions is the occurrence of chains of messages (segments) between two users (e.g. nodes 5 and 9 are likely to be posted by the same user). In consequence, some models define *segment lengths* as a feature or consider the authorship of messages to include *reciprocity* as another feature, i.e. users tend to reply to comments that replied to their previous messages. Furthermore, the consideration of authorship allows the definition of features related to *social influence* and *user roles*, i.e., certain users might follow a specific behavior. Some models also propose other types of features like the occurrence of certain text expressions.

5 Survey on generative models of online discussion threads

The models presented in this section are, to our best knowledge, the existing generative models of discussion threads in the state of the art:

- Kumar et al. [83]
- Wang et al. [84],
- Gómez et al. [85]
- Backstrom et al. [86]
- Nishi et al. [87]
- Lumbreras [88]
- Aragón et al. [89]

The selection of these models is based on the consideration of Kumar et al. [83] as the first generative model for online discussion threads. The rest of the models in the survey were selected after examining the publications citing this work in Scopus (52 papers) and Google Scholar (92 papers)[2], and including the studies which proposed a generative model for the structure and growth of online discussion threads.

To better identify the similarities and differences of the models of this survey, we present their main characteristics in Table 1. We observe the heterogeneity of these approaches in relation to features (popularity, novelty, reciprocity, root-bias, arrival patterns, text expressions, social influence, segment lengths, user roles), structure of threads (tree, array), temporal dimension (discrete, continuous), and the structural properties for validation (size, depth, degree, shapes). Also, the evaluation of the models is done with real data from online discussion platforms of very diverse nature: online forums (Y! Groups, Usenet), social news (Slashdot, Barrapunto, Digg, Reddit, Menéame), peer production (Wikipedia), social networks (Facebook, Google Plus), and microblogging services (Twitter).

Before presenting each generative model, we should introduce the Galton-Watson branching process for its history and relevance in modeling random trees [90]. Indeed, this model is often used as a baseline to compare against other models, e.g. [83]. It starts with a single root node, i.e. the post, and evolves at discrete time-steps. To generate the nodes at time-step $t + 1$, each node originated at time-step t generates independently a certain number of children deg according to a fixed probability distribution $p(deg)$. This process is repeated until no new children are generated, i.e. the discussion is over. This is a very simple model that can be estimated very efficiently from the data, since it just requires fitting the empirical distribution $p(deg)$. Although the classical branching process is able to reproduce certain features of online discussion threads such as the degree distribution, it is not a generative model that can explain the mechanisms underlying the dynamics of online discussions. Because, it uses a fixed probability distribution p at each node, it may fail to capture other relevant structural properties such as the depth distribution and it disregards the authorship and the arrival timestamp of the messages.

5.1 Kumar et al. [83]

The limitations of the classical branching process are addressed in Kumar et al. [83] by incorporating the novelty upon the preferential attachment model. That is to say, messages not only attract replies according to the number of previous replies, i.e. degree, but also to their time-stamp. At time-step t, either the thread terminates

Table 1 Main characteristics of the generative models of online discussion threads: the features, whether the predicted thread is a tree-like structure, whether threads grow in discrete or continuous time, and the datasets and structural properties used for the parameter estimation and the validation of the model

Model	Ref.	Features	Structure	Time	Datasets	Str. properties
Kumar et al.	[83]	Popularity, Novelty, Reciprocity	Tree	Discrete	Y! Groups, Usenet, Twitter	Size, Depth, Degree
Wang et al.	[84]	Popularity	Tree	Continuous	Digg, Reddit, Epinions	Size
Gómez et al.	[85]	Popularity, Novelty, Root-bias	Tree	Discrete	Slashdot, Barrapunto, Wikipedia, Menéame	Size, Depth, Degree
Backstrom et al.	[86]	Novelty, Arrival patterns, Text expressions, Social influence	Array	Continuous	Facebook, Google+, Wikipedia	Size
Nishi et al.	[87]	Popularity, Segment lengths	Tree	Discrete	Twitter	Size, Depth, Shapes
Lumbreras et al.	[88]	Popularity, Novelty, Root-bias, User Role	Tree	Discrete	Reddit	Size, Depth, Degree
Aragón et al.	[89]	Popularity, Novelty, Root-bias, Reciprocity	Tree	Discrete	Menéame	Size, Depth, Degree

with some fixed probability $p_f \in (0,1)$ or a new comment is attached to an existing comment k. At time t, the probability of attachment depends on two features: the popularity or degree $\deg_k$, and the novelty, or elapsed time since k was written, r_k. These features are parametrized by α and τ, respectively.

Formally, let the random discrete variable X_t denote the label of the parent node at time t. If a new node is attached, an existing node k, $k = 0, \ldots, t$, is chosen with probability proportional to a linear combination of the two previous features

$$p(X_t = k|\alpha, \tau, p_f) = \frac{\alpha \deg_k + \tau^{r_k}}{\sum_{k'} \left(\alpha \deg_{k'} + \tau^{r_{k'}}\right) + p_f}, \qquad (1)$$

where $\alpha \geq 0, \tau \in (0,1)$ and $p_f \in (0,1)$ are real numbers and the model parameters to be optimized for a given dataset.

Kumar et al. also propose an authorship model to determine the author of a comment based on the observation that users tend to reply users who had previous replied to their messages. In this model, the author of a new message is selected from the path between this new message and the root node with some probability, and otherwise randomly. However, this model is limited in the sense that the structure and growth of the discussion thread do not depend on the authorship of the messages.

5.2 Wang et al. [84]

An alternative framework for modeling the dynamics of online discussions is to consider a continuous-time model. This approach is more convenient when one is interested, for example, in understanding phenomena related to reaction times or lifespan of online conversations. Continuous-time models typically use counting processes as generative models [91]. Here, we focus on Wang et al. [84], in which commenting behavior is analyzed together with the topic (post) exposure duration to understand user attention to news items.

Their generative model is motivated by conflicting observations of previous studies that report significant differences in the probability distribution of thread sizes (the total number of comments of a conversation). While the threads sizes analyzed in some studies followed heavy-tailed distributions [83, 92], other previous studies reported distributions with a light tail [93, 94]. Wang et al. [84] first observes that the waiting time between two consecutive comments from a user follows a upper truncated Pareto distribution. Based on this observation, they proposed a model that explains the discrepancies of previous studies by means of the topic exposure duration distribution. The growth of attention eventually saturates because the old topics are replaced with newly generated contents.

One important assumption is that users share the same microscopic behaviors, i.e. the waiting time for different users comes from the same distribution. In doing so, they are able to model the process of M users as M independent concurrent counting process. For sites like Digg and Reddit with a short exposure distribution, the predicted distributions of sizes are also light-tailed, whereas other sites, like Epinions, with longer exposure durations, the obtained sizes are heavy tailed.

Wang et al. [84] focuses on reproducing the in-degree distribution of the comments. This is achieved by considering a preferential attachment process. In this model, t denotes the exact time passed since the creation of a topic. Let $\deg_k(t)$ denote the in-degree at time t for a comment k and let p_0 be the fixed probability to comment a

post/comment with no replies. The probability of a new comment attaching to comment k is given by

$$\frac{\deg_k(t) + p_0}{(1+p_0)\gamma M t^{c_0-c}} \tag{2}$$

for constants γ and c and a positive exponent c_0 that measures the combined impacts of factors such as resonance and social influence.

5.3 Gómez et al. [85]

This discrete-time model extends previous generative models for discussion threads [83, 92]. Besides popularity and novelty features (parameterized with α and τ respectively), it considers an additional feature (a root bias) that makes explicit the difference between the process of writing to the post (with id 0) node and to a descendant (user comment). This feature is parametrized with β, a positive real number.

Instead of having a parameter p_f to terminate the generation of a thread, as in Kumar et al. [83], this model generates threads of a given size, drawn from the empirical distribution of a dataset of conversation threads. Formally, the next parent node X_t is chosen according to the following probability:

$$p(X_t = k|\alpha,\tau,\beta) = \frac{\alpha\deg_k + \tau^{r_k} + \beta\delta_{0,k}}{\sum_{k'}\alpha\deg_{k'} + \tau^{r_{k'}} + \beta\delta_{0,k'}} \tag{3}$$

where $\delta_{0,k}$ is the Kronecker delta function, i.e. β is a free parameter for the root node, and zero otherwise. The relation between the model of Kumar et al. and this one is made clear by looking at both numerators of Eqs. (1) and (3).

In Gómez et al. [85], a model comparison was also done to show the relevance of each feature in every dataset. This statistical test was performed by considering the likelihoods of different reduced models that neglect each of the features separately.

5.4 Backstrom et al. [86]

The next model under consideration, Backstrom et al. [86], is not strictly speaking a generative model of discussion threads, but proposes how to predict structural properties of a thread (e.g. size) by combining features of different nature. In this model, the representation of discussion threads differs from the previous models. Here threads are represented as sequences of arrivals of comments regardless the reply relationship among them. This decision might be explained by the linear conversation view of the platforms used for the evaluation of the model, e.g. Facebook.

The model focuses on the authorship of the first comments of the sequence in order to predict, among other purposes, the final size of the thread. Each thread is represented with ρ, a sequence of non-negative integers in which the ρ_t is equal to the number of distinct users arriving to the discussion thread before the author of comment at time-step t ($\rho_t = 0$ if the author wrote the initial post). The data structure λ is then used to assess whether the five possible length-two patterns (0,0),(0,1),(1,0),(1,1),(1,2) have predictive value of the (macro-averaged) thread size. The predictive model of the thread size is then built using these arrival patterns along with some additional features:

- Social influence: Number of links between the user and users who previously commented, and number of links between the user and the user who published the post.
- Novelty: Elapsed (continuous) time for the first comments to be published.
- Text-based: The occurrence of terms like 'comment', 'agree', etc.
- Miscellany: Number of words, characters, and question/exclamation marks in the comment, and number of links in the post before and after the comment is posted.

5.5 Nishi et al. [87]

This model has been recently proposed for reply trees on Twitter. The model is motivated by observing that the structure of discussion threads is characterized by some long path-like reply trees, large star-like trees, and long irregular trees. Actually, some of the previous models already denoted long path-like reply trees as 'skinny' in Kumar et al. [83] or 'focused' in Backstrom et al. [86], and large star-like trees as 'bushy' in Kumar et al. [83] or 'expansionary' Backstrom et al. [86].

Because many of the previous models are based on the branching process model which does not capture appropriately long chains of messages in discussion threads, the depth distribution is often underestimated, as noted in Kumar et al. [83] and Gómez et al. [85]. This last model proves that the branching process model produces unrealistic fractions of long path-like trees or large irregular trees (combination of star-like and path-like structures). Therefore, the authors introduce the concept of segments: maximal chains without branching in a discussion thread. Formally, a segment of length λ is defined by $\lambda + 1$ connected nodes (replies) such that the $\lambda - 1$ inner nodes have in-degree equals to 1. For example, the discussion thread in Fig. 5 is composed by 5 segments of $\lambda = 1$ ($1-2$, $1-3$, $1-8$, $2-4$, $2-7$) and a segment of $\lambda = 3$ ($1-5-6-9$). Thus, the model adds to the branching process model: (1) the distribution of segment length l, (2) the correlation between λ and the degree of the root, and (3) the correlation between the degree of root and the the degree of the end node of segments.

The assessment of this extension shows the ability to capture the fraction of long path-like trees but not large irregular trees. According to the authors, results are explained because a large λ value in one branch implies a relatively high probability of large λ values in other branches. This effect is solved by a final extension which allows λ to be correlated among segments starting from the same node.

5.6 Lumbreras [88]

This generative model is part of a doctoral thesis about automatic role detection in online forums [88]. It is motivated by observing that the growth of discussion threads in previous generative models, in particular Kumar et al. [83] and Gómez et al. [85], is irrespective of the user who is writing a new message.

This new model proposes that there might exist different roles which categorize users who participate in the discussion threads. To this end, the model builds upon Gómez et al. [85] and introduces latent types of users or roles. In this model, a role u, $u = 1, \ldots, U$, corresponds to specific values $\theta_u = (\alpha_u, \beta_u, \tau_u)$ associated to the popularity, root-bias and novelty influence, respectively, of a type of user.

Let z_k denote a binary vector of U entries indicating the role membership of the author of the k-th comment, i.e., $z_{ku} = 1$ if author of comment k belongs to role u, and zero otherwise. This is the latent variable not present in the data. Let q_u denote its marginal distribution, i.e., $p(z_{ku} = 1) = q_u$ with $q_u \geq 0$ and $\sum_{u=1}^{U} q_u = 1$.

In this model, the next parent node X_t is chosen according to the following joint probability:

$$p(X_t = k, z_k|\theta) = \prod_{u=1}^{U} q_u^{z_{ku}} p(X_t = k|\theta_u)^{z_{ku}}, \tag{4}$$

where $p(X_t = k|\theta_u)$ is the same as Eq. (3).

The existence of the latent variables z prevents to optimize a complete likelihood function defined using Eq. (4). Therefore, the expectation-maximization algorithm is used as an optimization procedure. The number of roles U (model selection) is computed using Bayesian Information Criteria.

Roles are finally used to analyze their predictive power, i.e. the capability of this model to predict the parent message of arriving messages in comparison to Gómez et al. [85], and two minimal models: one based on popularity [81] and the other on novelty.

5.7 Aragón et al. [89]

The last generative model of this survey considers the comments authorships in a novel way. In Aragón et al. [89], both authorships and thread structure co-evolve simultaneously, and mutually depend one each other during the evolution of the conversation. The model maintains two chains, one for the authors and one for the comments. At time t, the state of a discussion is given by a vector of authors ids $a_{1:t}$ and a vector of parents $f_{1:t}$.

On the one hand, the authorship evolves according to a preferential attachment process. With probability p_{new} (estimated from the data), a new author joins the discussion and, otherwise, an existing author ν is chosen with probability that depends exponentially on the number R_ν of times ν has been replied in the thread:

$$p(a_{t+1} = \nu|a_{1:t}, \pi_{1:t}) = \begin{cases} p_{new}, & \text{for } \nu = V + 1 \\ \frac{(1-p_{new})2^{R_\nu}}{\sum_{i=1}^{U} 2^{R_i}}, & \text{for } \nu \in 1, \ldots, V \end{cases} \tag{5}$$

where V is the number of different users in the discussion thread so far.

On the other hand, the thread structure evolves as in Gómez et al. [85], with an additional feature $\delta_{a_{t+1}, a_{\pi_k}}$ representing the author reciprocity of a message, i.e. whether the selected user a_{t+1} is the author of the parent message a_{π_k}:

$$\begin{aligned} \phi_k(\pi_{1:t}, a_{1:t}; \theta) &= \alpha \deg_k + \tau^{r_k} + \beta \delta_{\text{root},k} + \kappa \delta_{a_{\pi_k}, a_{t+1}} \\ p(\pi_{t+1} = k|\pi_{1:t}, a_{1:t}; \theta) &= \frac{\phi_k(\pi_{1:t}, a_{1:t}; \theta)}{\sum_{k'} \phi_{k'}(\pi_{1:t}, a_{1:t}; \theta)}, \end{aligned} \tag{6}$$

for parameters $\theta = (\alpha, \beta, \tau, \kappa)$. The likelihood optimization in this model is more expensive than for the previous ones, since the normalization in (6), contrary to the one in Eq. 3, does not only depend on time, but also on the structure of the thread.

6 Applications of modeling the structure of online discussions

The development of generative models and their statistical assessment with empirical data aim to characterize the mechanisms governing the dynamics of online discussion. Moreover, each model from the previous section was also motivated by specific research questions which were addressed by selecting features carefully and using data from specific online discussion platforms (see an overview in Table 1). Thus, these particular objectives reveal some practical applications of modeling the structure of online discussion. In this section we present how the selected generative models were used to compare dynamics of online discussion among different platforms, to predict some patterns of user behavior, and to evaluate the impact of design on online discussion platforms. We also sketch the general potential for applications of generative models in these different application areas.

6.1 Comparison among online discussion platforms

The most straightforward application of generative models for online discussion threads is the interpretation of its parameters, i.e. the quantification of the relevance of each model feature, to compare platforms of different nature.. This is a first step towards a more fine grained analysis to estimate the impact of different platform design elements on how users engage in written online discussions and will be explored more detail in Section 6.3 and has a great potential impact in assessing user interface design choices. It can be used as well on the same underlying platform to compare different user communities but for example in different language versions or different spheres of interest, which allows then to measure the impact of a specific topic or cultural aspect.

Four of the generative models surveyed in this article: Kumar et al. [83], Wang et al. [84], Gómez et al. [85], and Backstrom et al. [86] were validated in this respect with data from multiple online discussion platforms.

The approach in Kumar et al. [83] was validated with data from Y! Groups, Usenet groups, and Twitter. While the results with data from Y! Groups was not very informative, results with discussion threads from Usenet revealed that political groups exhibit greater degree of preferential attachment (popularity), groups with fewer users are more affected by novelty, and less new authors tend to join Q&A groups. In the case of Twitter, the comparison among discussions around different hashtags served to discover that novelty is prominent in threads about topics with a stronger sense of time, e.g. sports.

The validation of the model in Wang et al. [84] relied on data from two social news sites (Digg and Reddit) and a consumer review site (Epinions). Interestingly, this model emphasizes the ability to characterize the heterogeneity of the size distribution of conversations across these three platforms. In particular, results reveal that the size distribution of discussions in Digg and Reddit is light-tailed but heavy-tailed in Epinions. This is not the case for the in-degree distribution of comments, which follows a Pareto distribution in the three platforms.

The model in Gómez et al. [85] was validated with data from three social news sites (Slashdot, Barrapunto, Menéame) and the talk pages from Wikipedia. Results reveal that popularity is important in the social news sites but, in contrast, irrelevant in the growth of Wikipedia discussions. Moreover, the root-bias presents a much stronger relevance in Menéame, a platform which displayed the comments linearly regardless of the reply relationship. Another interesting aspect of Gómez et al. [85] was its comparison of Slashdot and Barrapunto, which used the same underlying platform but were run in two different languages (English and Spanish) which resulted in two very distinctive user communities. The modeling approach revealed the larger impact of novelty in the Spanish community while popularity was more important in Slashdot.

The approach in Backstrom et al. [86] relied on data from Facebook and Wikipedia. The validation precisely focused on distinguishing which features are key to understand the size of discussion threads. In both platforms, the elapsed time between and the last comment from the early sequence of comments becomes very informative about the size of the final thread. However, the most relevant feature in Wikipedia is the length of the last comment. Therefore, although discussions in both platforms might be explained with time-based features, content-based features are even stronger in a peer production online environment as Wikipedia.

6.2 Prediction of user behavior

Features in models for the structure of online discussion can also serve to predict behavior. This is the case of Backstrom et al. [86] which, as discussed above, is not strictly speaking a generative model of discussion threads but a predictive model. A first analysis of the empirical shows that threads are longer when (1) the first comment authors are friends, and (2) the elapsed time between the publication of the post and the first comment is lower. This motivates the definition of the predictive model which considers early sequences of threads to infer the final thread size. Results confirm that the five possible length-two patterns of authors commenting do have predictive value of the (macro-averaged) thread size. Then, a broad range of features is proposed, including features extracted via text regression. As one could expect, the combination of all features exhibits the best performance, although text-regression features sometimes perform worse than a pseudo-random baseline.

Lumbreras [88] is the second model of our selection which also addresses prediction of user behavior. In particular, the model examines whether the identification of groups of users with the same behavior (i.e identical parameter values of popularity, novelty and root-bias) might predict user behavior in a new context. The model is validated with different datasets from Reddit to confirm that the approach is able to detect different number and types of user behavior. That is to say that there are different user roles which describe how users participate differently in online discussion threads. However, the validation of the predictive model shows that the predictive power of these roles is almost marginal.

Although the other studies presented in the previous section have not explored these predictive capabilities, all of these generative models have inherently the capability to be used to predict the future evolution of an online discussion given its state at a given point in time. It is thus an interesting topic for future research in particular through adding more temporal or user based features.

6.3 Evaluation of platform design

The evaluation of platform design elements is probably the most useful application from a technical point of view as it can be used to asses the impact of a given design element on the user interaction patterns on a platform.

The suitability of this approach has been shown by Aragón et al. [89], the most recent generative model of the survey, which was motivated by a very a specific research question about platform design. This dealt with the change of how conversation threads are presented in the social news site Menéame. Threads were originally presented in a linear view regardless of the reply relationship and sorted cronologically. Indeed, this was the interface when the data for Gómez et al. [85] was retrieved, concluding that the root-bias was much stronger in Menéame than in the other platforms which presented threads hierarchically in a tree-like structure. However, Menéame replaced the original linear view for a hierarchical view in 2015.

Given that threads in online discussion are commonly characterized by long chains of reciprocal of messages (visually emphasized in hierarchical views) [95], Aragón et al. [89] incorporates reciprocity to the original features from Gómez et al. [85]: popularity, root-bias and novelty. Results before and after the change confirm that the new hierarchical design induced more reciprocal activity, made popular comments to attract more replies and slowed down the decay of novelty. This finding shows the huge potential of the generative modeling approach to help to assess the interdependency between users interaction patterns and platform design elements. This can be exploited to help site owners and community managers to create a positive and constructive environment for large scale online discussions.

7 Open research challenges

We have presented several models that are able to reproduce many of the characteristics of online discussion threads in platforms of very different nature. By observing the state of the art models, we have identified some relevant issues which have not been addressed yet and are expected to receive growing attention in the following years.

7.1 Competition between discussion threads

The generative models described above are able to reproduce patterns of user commenting behavior. However, there is a lack of understanding of the key factors that determine why a given user will write a comment in a particular thread and not on another. Specifically: to what extent a user comment is determined by the opinion of other users, the topic of the news post, or how popular or recent the post is? More generally, is there a global mechanism that can capture how the messages of different users distribute themselves among the different posts available? And what are the identifiable features of that mechanism? Some studies on Digg [96, 97] explored how threads receive incoming votes over time, even, before and after being selected for the front page. However, there are no models which explain (1) how threads compete among themselves to receive attraction from users and/or (2) which features are the most appealing to users when posting a message in a pool of candidate threads.

The models selected in this survey only characterize the arrival of comments to a single thread. The only exception to this is Backstrom et al. [86] in which the thread to be commented is picked from a dynamic list of threads. However, the model establishes a fixed probability to every thread. In turn, it might be of interest that arriving comments should be able to reply comments from a set of threads and, furthermore, arriving nodes could also be the initial nodes of new discussion threads. Thus, instead of setting a list of equally available threads as done in Wang et al. [84], an extended model could estimate the longevity of discussion threads to minimize the arrival of new comments to old conversations with little interest within the online community. An alternative approach that might be explored is modeling this research challenge as a competition between conversation threads when bringing attraction of users [98]. In particular, the model considers that although arriving nodes are likely to link high-connected nodes, making connections to those nodes should be more expensive. That is to say, competition is conceived as a tradeoff between connectivity and cost. In this way, future work should explore that users are interested in replying comments in popular debates but the emergence of new discussion threads might reduce their exposure and, therefore, the cost of getting access to the old ones should be higher.

7.2 Groups of users

We have observed that homophily and social influence are features of online interaction that usually induce a segregation or clustering in the community [47]. Although some models (e.g. Backstrom et al. [86]) include social influence as a formal feature, none of them include homophily. Such extension is far from trivial because it would require to model the existence of groups of users, with common interests or similar opinion about certain topics. This open research challenge is relevant because user groups usually evolve into echo chambers [30], which might favor extremism. This leads to us to reflect on how generative models would better explain online discussion if groups of users were taken into account.

In relation to this issue, two of the main research challenges in the topic of homophily are multiplexity, i.e. understanding the role of networks with various layers of

interaction types, and analyzing dynamic data in which links appear and disappear over time [16]. Some studies reviewed in this article showed that homophily is stronger in positive interactions rather than in reply interactions [37]. Thus, the problem of identifying groups of users can be viewed as a community detection problem. This could be handled by the many existing algorithms [99] on the network of votes among users within the discussion. On the other hand, this could also motivate new methods to detect communities based on interactions (i.e. comments or votes) which only occur between opposing fractions. In addition, the co-evolution of votes and comments is currently receiving increasing attention [100]. Therefore, we suggest that the research challenge of modeling online discussions including groups of users (to be defined with interactions from the voting layer) will provide a better explanation of the behavior of online communities in discussion platforms.

7.3 The role of content

Previous research, reviewed in Section 3.3, indicated that emotional contagion in online discussions is a very present phenomenon. For instance, emotional expressions prolong online discussions [67]. In contrast, most of the generative models reviewed in this article do not include features related to the content of messages within the discussion, with the only exception of Backstrom et al. [86] which consider some text-based features like the occurrence of certain terms (e.g. 'comment', 'agree') or the number of question/exclamation marks in a comment. We should note that language-independent approaches are easily replicable in online discussions of very diverse nature. However, we also consider that only focusing on structural aspects of threads might be limiting when characterizing online discussion.

Content-based approaches are mandatory for relevant research topics like modeling trolling behavior [101, 102]. Because modeling online discussion relies on representing discussion threads as information cascades, generative models could be enriched with existing methodologies of emotional cascades of online activity [70]. Therefore, understanding collective emotions in online discussion is still a challenging task, requiring generative mechanisms that can bridge individual and collective levels of behavior [103].

Besides emotions, the content of messages can also reveal the emergence and evolution of topics in online discussions. For instance, some studies have found strong evidence that hierarchical comment threads represent a topical hierarchy in discussion platforms [104]. Thus, this observation explicitly motivates the inclusion of text-based features (e.g. text similarity between replies) to better characterize the arrival of new comments in a discussion thread. To our best knowledge, this approach has never been considered in any generative model and, therefore, examining the role of content is still a open research challenge in modeling online discussion.

7.4 Influencing user activity

An important challenge nowadays is how to devise strategies to influence, or reshape, user activity. This type of problem has been traditionally formulated as a social influence problem, in which a set of users needs to be found in order to maximize, for example, the impact of a cascade of product adoptions [105]. A perhaps more principled approach could be to learn a policy or control law that guides the user activities in a closed-loop setting, where user feedbacks are incorporated during the cascade.

This problem has started to be addressed from areas such as optimal control or reinforcement learning [106]. Most works so far formulate continuous-time models using temporal point processes and are focused on social influence, such as guiding opinion diffusion [107] or determining when to post to maximize impact [108]. The proposed models in this review could be used, for example, to learn a platform dynamics of commenting behavior which is then influenced using some control mechanism on the platform, as has been recently proposed [109].

It remains an open question whether these methods, which are computationally demanding and are limited by their model assumptions, can be deployed effectively in real platforms.

8 Conclusion

Online discussion threads have received increasing interest from academia in order to provide a better understanding of the principles of human communication. The ease of extracting discussion threads from online platforms has promoted the development of statistical models which have been proven effective to validate theories of user and social behavior. In this paper, we have surveyed the state of the art in modeling the structure of online discussion, including a historical overview, empirical evidence of relevant social theories, and the description and applications of seven statistical models to reproduce the structure and growth of discussion threads.

Despite the notable findings from these models, some important issues remain unaddressed in this domain. In particular, we have found of interest to explore the competition between discussion threads, the existence of groups of users and the role of content in the formation of online conversations, and how to influence user activity. Therefore, we believe that these research gaps become an excellent opportunity to improve the characterization of

online discussion. Because online discussions in the form of written comments are expected to remain popular over the following years and play a key role in relevant processes like the formation public opinion, online education systems, peer production environments, and civic participation for policy making, we aim this work to be helpful to identify open research challenges and to motivate future work in modeling online discussions.

Endnotes

[1] http://www.4chan.org/advertise

[2] The number of citations when this survey was done.

Abbreviations

BBS: Bulletin board systems; IRC: Internet relay chat; Q&A: Question and answer

Acknowledgments

We would like to thank David Laniado for his valuable discussions and suggestions that helped to improve the manuscript.

Funding

This work has been supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502); the Marie Curie FP7-PEOPLE-2012-COFUND Action (grant agreement no: 600387); and the CIEN LPS-BIGGER project (UCTR150175, IDI-20141259), co-funded by Centro para el Desarrollo Tecnológico Industrial (CDTI) and Fondo Europeo de Desarrollo Regional (FEDER). The funders had no role in study design, decision to publish, or preparation of the manuscript.

Availability of data and materials

N/A. No data was used for this survey, all references are provided.

Authors' contributions

PA, VG, DG and AK made a substantial contribution to this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that no competing interests exist.

Publisher's Note

Author details

[1]Universitat Pompeu Fabra, Barcelona, Spain. [2]Eurecat - Technology Centre of Catalonia, Barcelona, Spain. [3]ETH Zürich, Zurich, Switzerland. [4]NTENT, Av. Diagonal 220, 08018 Barcelona, Spain.

Received: 16 January 2017 Accepted: 27 September 2017
Published online: 05 October 2017

References

1. Kemp S. Digital, social & mobile worldwide in 2016. We are social. 2016. https://wearesocial.com/uk/special-reports/digital-in-2016.
2. Rossi LA, Gnawali O. Language independent analysis and classification of discussion threads in Coursera MOOC forums. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014). IEEE; 2014. p. 654–61. doi:10.1109/IRI.2014.7051952.
3. Qu Y, Wu PF, Wang X. Online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake. In: System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on. IEEE; 2009. p. 1–11.
4. Zubiaga A, Liakata M, Procter R, Hoi GWS, Tolmie P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE. 2016;11(3):150989.
5. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, et al. Life in the network: the coming age of computational social science. Science (New York, NY). 2009;323(5915):721.
6. Anderson C. The end of theory: The data deluge makes the scientific method obsolete. Wired Mag. 2008;16(7):16–07.
7. Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM, et al. A survey of statistical network models. Foundations Trends® Mach Learn. 2010;2(2): 129–233.
8. Lusher D, Koskinen J, Robins G. Exponential random graph models for social networks: theory, methods, and applications. In: Structural Analysis in the Social Sciences. Cambridge: Cambridge University Press; 2012. doi:10.1017/CBO9780511894701. https://www.cambridge.org/core/books/exponential-random-graph-models-for-socialnetworks/9296EE2B53CDEF9FE9E2E981E2FDB8A8.
9. Calvão L, Pimentel M, Fuks H. Internet conversation media: an evolutionary perspective from email to social networks. Rio de Janeiro: UNIRIO; 2016.
10. Wooley DR. Talkomatic program. Urbana-Champaign: University of Illinois Urbana-Champaign; 1972.
11. Oikarinen J. Reed D. Internet relay chat protocol. 1988. https://buildbot.tools.ietf.org/html/rfc1459.
12. Doc G. "28/85" "services and facilities to be provided in the gsm system" rev2. 1985. http://www.etsi.org/deliver/etsi_gts/01/0102/05.00.00_60/gsmts_0102v050000p.pdf.
13. Daniel S, Ellis J, Truscott T. Usenet, a general access unix network. Durham: Duke University; 1980.
14. Berners-Lee T. Information management: A proposal. 1989. http://www.etsi.org/deliver/etsi_gts/01/0102/05.00.00_60/gsmts_0102v050000p.pdf.
15. O'Reilly T. Web 2.0: compact definition. 2005. http://radar.oreilly.com/2005/10/web-20-compact-definition.html.
16. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annu Rev Sociol. 2001;27(1):415–44. doi:10.1146/annurev.soc.27.1.415. https://doi.org/10.1146/annurev.soc.27.1.415.
17. Lazarsfeld PF, Merton RK, et al. Friendship as a social process: A substantive and methodological analysis. Freedom Control Modern Soc. 1954;18(1):18–66.
18. Newman ME. Mixing patterns in networks. Phys Rev E. 2003;67(2):026126.
19. Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. Stat Med. 2013;32(4):556–77.
20. Singla P, Richardson M. Yes, there is a correlation:-from social networks to personal behavior on the web. In: Proceedings of the 17th international conference on World Wide Web. New York: ACM; 2008. p. 655–64. doi:10.1145/1367497.1367586. http://doi.acm.org/10.1145/1367497.1367586.
21. Thelwall M. Homophily in myspace. J Am Soc Inf Sci Technol. 2009;60(2): 219–31.
22. Jernigan C, Gaydar MistreeBF. Facebook friendships expose sexual orientation. First Monday. 2009;14(10):2009.
23. Sarigol E, Garcia D, Schweitzer F. Online privacy as a collective phenomenon. In: Proceedings of the second ACM conference on Online social networks. New York: ACM; 2014. p. 95–106. doi:10.1145/2660460.2660470. http://doi.acm.org/10.1145/2660460.2660470.
24. Laniado D, Volkovich Y, Kappler K, Kaltenbrunner A. Gender homophily in online dyadic and triadic relationships. EPJ Data Sci. 2016;5(19):19.
25. Lönnqvist J-E, Itkonen JV. Homogeneity of personal values and personality traits in facebook social networks. J Res Persona. 2016;60: 24–35.
26. Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. Bmj. 2008;a2338:337.
27. Bliss CA, Kloumann IM, Harris KD, Danforth CM, Dodds PS. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. J Comput Sci. 2012;3:388–97.
28. Bollen J, Gonċalves B, Ruan G, Mao H. Happiness is assortative in online social networks. Artif Life. 2011;17(3):237–51.
29. Adamic LA, Glance N. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In: Proceedings of the 3rd International

Workshop on Link Discovery. New York: ACM; 2005. p. 36–43. doi:10.1145/1134271.1134277. http://doi.acm.org/10.1145/1134271.1134277.
30. Sunstein CR. Republic. com 2.0. Princeton: Princeton University Press; 2007.
31. Pariser E. The filter bubble: What the Internet is hiding from you. London: Penguin; 2011.
32. Bakshy E, Messing S, Adamic LA. Exposure to ideologically diverse news and opinion on facebook. Science. 2015;348(6239):1130–2.
33. Barberá P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. Political Anal. 2015;23(1):76–91.
34. Conover M, Ratkiewicz J, Francisco MR, Gonċalves B, Menczer F, Flammini A. Political polarization on twitter. ICWSM. 2011;133:89–96.
35. Lietz H, Wagner C, Bleier A, Strohmaier M. When politicians talk: Assessing online conversational practices of political parties on twitter. Ann Arbor: The AAAI Press; 2014.
36. Aragón P, Kappler KE, Kaltenbrunner A, Laniado D, Volkovich Y. Communication dynamics in twitter during political campaigns: The case of the 2011 spanish national election. Policy Internet. 2013;5(2):183–206.
37. Garcia D, Abisheva A, Schweighofer S, Serdült U, Schweitzer F. Ideological and temporal components of network polarization in online political participatory media. Policy & Internet. 2015;7(1):46–79.
38. Neff JJ, Laniado D, Kappler KE, Volkovich Y, Aragón P, Kaltenbrunner A. Jointly they edit: Examining the impact of community identification on political interaction in wikipedia. PLoS ONE. 2013;8(4):e60584.
39. Laniado D, Tasso R, Volkovich Y, Kaltenbrunner A. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In: ICWSM-11 - 5th International AAAI Conference on Weblogs and Social Media. Barcelona: The AAAI Press; 2011.
40. Laniado D, Tasso R. Co-authorship 2.0: Patterns of Collaboration in Wikipedia. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. New York: ACM; 2011. p. 201–10. doi:10.1145/1995966.1995994. http://doi.acm.org/10.1145/1995966.1995994.
41. Deutsch M, Gerard HB. A study of normative and informational social influences upon individual judgment. J Abnorm Soc Psychol. 1955;51(3):629.
42. Kelman HC. Compliance, identification, and internalization three processes of attitude change. J Confl Resolut. 1958;2(1):51–60. doi:10.1177/002200275800200106. https://doi.org/10.1177/002200275800200106.
43. Granovetter M. Threshold models of collective behavior. Am J Sociol. 1978;83(6):1420–43. doi:10.1086/226707. https://doi.org/10.1086/226707.
44. Watts DJ, Dodds PS. Influentials, networks, and public opinion formation. J Consum Res. 2007;34(4):441–58.
45. Easley D, Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge: Cambridge University Press; 2010.
46. Quattrociocchi W, Caldarelli G, Scala A. Opinion dynamics on interacting networks: media competition and social influence. Sci Rep. 2014;4. doi:10.1038/srep04938. http://dx.doi.org/10.1038/srep04938.
47. Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Natl Acad Sci. 2009;106(51):21544–9.
48. Cha M, Mislove A, Gummadi KP. A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th International Conference on World Wide Web. New York: ACM; 2009. p. 721–30. doi:10.1145/1526709.1526806. http://doi.acm.org/10.1145/1526709.1526806.
49. Ye S, Wu SF. Measuring message propagation and social influence on twitter.com. In: Proceedings of the Second International Conference on Social Informatics. Berlin, Heidelberg: Springer-Verlag; 2010. p. 216–31. http://dl.acm.org/citation.cfm?id=1929326.1929342.
50. Dow PA, Adamic LA, Friggeri A. The anatomy of large facebook cascades. In: ICWSM-13 - 7th International AAAI Conference on Weblogs and Social Media. Boston: The AAAI Press; 2013.
51. Dholakia UM, Bagozzi RP, Pearo LK. A social influence model of consumer participation in network-and small-group-based virtual communities. Int J Res Mark. 2004;21(3):241–63.
52. Cho H, Stefanone M, Gay G. Social information sharing in a CSCL Community. In: Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community. Boulder: International Society of the Learning Sciences; 2002. p. 43–50. http://dl.acm.org/citation.cfm?id=1658616.1658623.
53. Hsu C-L, Lin JC-C. Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. Inf Manag. 2008;45(1):65–74.
54. Huffaker D. Dimensions of leadership and social influence in online communities. Hum Commun Res. 2010;36(4):593–617.
55. Tan C, Tang J, Sun J, Lin Q, Wang F. Social action tracking via noise tolerant time-varying factor graphs. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2010. p. 1049–58. doi:10.1145/1835804.1835936. http://doi.acm.org/10.1145/1835804.1835936.
56. Walther JB, DeAndrea D, Kim J, Anthony JC. The influence of online comments on perceptions of antimarijuana public service announcements on youtube. Hum Commun Res. 2010;36(4):469–92.
57. Hatfield E, Cacioppo JT, Rapson RL. Emotional contagion. Cambridge: Cambridge University Press; 1994.
58. Hatfield E, Cacioppo JT, Rapson RL. Emotional contagion. Curr Dir Psychol Sci. 1993;2(3):96–100.
59. Kappas A. Social regulation of emotion: messy layers. Front Psychol. 2013;4:51.
60. Derks D, Fischer AH, Bos AE. The role of emotion in computer-mediated communication: A review. Comput Hum Behav. 2008;24(3):766–85.
61. Garcia D, Kappas A, Küster D, Schweitzer F. The dynamics of emotions in online interaction. Open Sci. 2016;3:8.
62. Coviello L, Sohn Y, Kramer ADI, Marlow C, Franceschetti M, Christakis NA, Fowler JH. Detecting emotional contagion in massive social networks. PLOS ONE. 2014;9:1–6.
63. Kramer AD, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci. 2014;111(24):8788–90.
64. Ferrara E, Yang Z. Measuring emotional contagion in social media. PLoS ONE. 2015;10:1–14.
65. Fan R, Zhao J, Chen Y, Xu K. Anger is more influential than joy: Sentiment correlation in weibo. PLoS ONE. 2014;9:1–8.
66. Iosub D, Laniado D, Castillo C, Fuster Morell M, Kaltenbrunner A. Emotions under discussion: Gender, status and communication in online collaboration. PLOS ONE. 2014;9:1–23.
67. Chmiel A, Sobkowicz P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M, Hołyst JA. Negative emotions boost user activity at bbc forum. Physica A: Stat Mech Appl. 2011;390(16):2936–44.
68. Pfitzner R, Garas A, Schweitzer F. Emotional divergence influences information spreading in twitter. ICWSM. 2012;12:2–5.
69. Fan R, Xu K, Zhao J. Higher contagion and weaker ties mean anger spreads faster than joy in social media. ArXiv e-prints. 2016. Provided by the SAO/NASA Astrophysics Data System. http://adsabs.harvard.edu/abs/2016arXiv160803656F.
70. Alvarez R, Garcia D, Moreno Y, Schweitzer F. Sentiment cascades in the 15m movement. EPJ Data Sci. 2015;4(1):1–13.
71. von Scheve C, Salmella M. Collective emotions. Oxford: Oxford University Press; 2014.
72. Chmiel A, Sienkiewicz J, Thelwall M, Paltoglou G, Buckley K, Kappas A. Hołyst JA. Collective emotions online and their influence on community life. PLoS ONE. 2011;6(7):e22207.
73. Garas A, Garcia D, Skowron M, Schweitzer F. Emotional persistence in online chatting communities. Sci Rep. 2012;2:402.
74. Sprent P. Data driven statistical methods: Taylor & Francis; 1997. https://books.google.es/books?id=IgziRAAACAAJ.
75. Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press; 2012.
76. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. Internet Math. 2004;1(2):226–51.
77. Browne MW, Cudeck R, et al. Alternative ways of assessing model fit. Sage Focus Editions. 1993;154:136.
78. Kolmogorov A. Sulla determinazione empirica di una lgge di distribuzione. Inst. Ital. Attuari, Giorn. 1933;4:83–91.
79. Smirnov N. Table for estimating the goodness of fit of empirical distributions. Ann Math Stat. 1948;19(2):279–81.
80. Arnold TB, Emerson JW. Nonparametric goodness-of-fit tests for discrete null distributions. R Journal. 2011;3(2).

81. Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
82. Simon HA. On a class of skew distribution functions. Biometrika. 1955;42(3/4):425–40.
83. Kumar R, Mahdian M, McGlohon M. Dynamics of conversations. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2010. p. 553–62. doi:10.1145/1835804.1835875. http://doi.acm.org/10.1145/1835804.1835875.
84. Wang C, Ye M, Huberman BA. From user comments to on-line conversations. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2012. p. 244–52. doi:10.1145/2339530.2339573. http://doi.acm.org/10.1145/2339530.2339573.
85. Gómez V, Kappen HJ, Litvak N, Kaltenbrunner A. A likelihood-based framework for the analysis of discussion threads. World Wide Web. 2013;16(5-6):645–75.
86. Backstrom L, Kleinberg J, Lee L, Danescu-Niculescu-Mizil C. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In: Proceedings of the sixth ACM international conference on Web search and data mining. New York: ACM; 2013. p. 13–22. doi:10.1145/2433396.2433401. http://doi.acm.org/10.1145/2433396.2433401.
87. Nishi R, Takaguchi T, Oka K, Maehara T, Toyoda M, Kawarabayashi K-I, Masuda N. Reply trees in twitter: data analysis and branching process models. Soc Netw Anal Mining. 2016;6(1):1–13.
88. Lumbreras A. Automatic role detection in online forums. PhD thesis, Ecole doctorale de InfoMaths (ED 512) LYON. 2016.
89. Aragón P, Gómez V, Kaltenbrunner A. To thread or not to thread: The impact of conversation threading on online discussion. Montreal: The AAAI Press; 2017.
90. Watson HW, Galton F. On the probability of the extinction of families. J Anthropol Inst G B Irel. 1875;4:138–44.
91. Daley DJ, Vere-Jones D. An introduction to the theory of point processes. vol. I., Elementary theory and methods. Probability and its applications. New York, Berlin, Paris: Springer; 2003.
92. Gómez V, Kappen HJ, Kaltenbrunner A. Modeling the structure and evolution of discussion cascades. In: Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia. New York: ACM; 2011. p. 181–90. doi:10.1145/1995966.1995992. http://doi.acm.org/10.1145/1995966.1995992.
93. Ogilvie P. Modeling blog post comment counts. 2008. URL http://livewebir.com/blog/2008/07/modeling-blog-post-comment-counts. Accessed 24 Nov 2010.
94. Tsagkias M, Weerkamp W, De Rijke M. Predicting the volume of comments on online news stories. In: Proceedings of the 18th ACM conference on Information and knowledge management. New York: ACM; 2009. p. 1765–8. doi:10.1145/1645953.1646225. http://doi.acm.org/10.1145/1645953.1646225.
95. Aragón P, Gómez V, Kaltenbrunner A. Detecting platform effects in online discussions. Policy Internet. doi:10.1002/poi3.158. http://dx.doi.org/10.1002/poi3.158.
96. Lerman K, Galstyan A. Analysis of social voting patterns on digg. In: Proceedings of the first workshop on Online social networks. New York: ACM; 2008. p. 7–12. doi:10.1145/1397735.1397738. http://doi.acm.org/10.1145/1397735.1397738.
97. Lerman K, Hogg T. Using a model of social dynamics to predict popularity of news. In: Proceedings of the 19th international conference on World wide web. New York: ACM; 2010. p. 621–30. doi:10.1145/1772690.1772754. http://doi.acm.org/10.1145/1772690.1772754.
98. D'souza RM, Borgs C, Chayes JT, Berger N, Kleinberg RD. Emergence of tempered preferential attachment from optimization. Proc Natl Acad Sci. 2007;104(15):6112–7.
99. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3):75–174.
100. Costa AF, Traina AJM, Traina C, Faloutsos C. Vote-and-comment: Modeling the coevolution of user interactions in social voting web sites. In: 2016 IEEE 16th International Conference on Data Mining (ICDM); 2016. p. 91–100. doi:10.1109/ICDM.2016.0020.
101. Cheng J, Danescu-Niculescu-Mizil C, Leskovec J. Antisocial behavior in online discussion communities. In: ICWSM-15 - 9th International AAAI Conference on Weblogs and Social Media. Oxford: The AAAI Press; 2015.
102. Cheng J, Bernstein M, Danescu-Niculescu-Mizil C, Leskovec J. Anyone can become a troll: Causes of trolling behavior in online discussions. New York: ACM; 2017. p. 1217–30. doi:10.1145/2998181.2998213. http://doi.acm.org/10.1145/2998181.2998213.
103. Schweitzer F, Garcia D. An agent-based model of collective emotions in online communities. Eur Phys JB. 2010;77(4):533–45.
104. Weninger T, Zhu XA, Han J. An exploration of discussion threads in social news sites: A case study of the reddit community. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM; 2013. p. 579–83. doi:10.1145/2492517.2492646. http://doi.acm.org/10.1145/2492517.2492646.
105. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2003. p. 137–46. doi:10.1145/956750.956769. http://doi.acm.org/10.1145/956750.956769.
106. Sutton RS, Barto AG. Introduction to Reinforcement Learning, 1st. Cambridge: MIT Press; 1998.
107. Wang Y, Williams G, Theodorou E, Song L. Variational policy for guiding point processes. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. International Convention Centre, Sydney: PMLR; 2017. p. 3684–93. http://proceedings.mlr.press/v70/wang17k.html.
108. Zarezade A, Upadhyay U, Rabiee HR, Gomez-Rodriguez M. Redqueen: An online algorithm for smart broadcasting in social networks. New York: ACM; 2017, pp. 51–60. doi:10.1145/3018661.3018684. http://doi.acm.org/10.1145/3018661.3018684.
109. Thalmeier D, Gómez V, Kappen HJ. Action selection in growing state spaces: control of network structure growth. J Phys A Math Theor. 2017;50(3):034006.