Large-scale forecasting of information spreading

Severiukhina, O.; Kesarev, S.; Bochenina, K.; Boukhanovsky, A.; Lees, M.H.; Sloot, P.M.A.

DOI
10.1186/s40537-020-00350-5

Publication date
2020

Document Version
Final published version

Published in
Journal of Big Data

License
CC BY

Link to publication

Citation for published version (APA):
Severiukhina, O., Kesarev, S., Bochenina, K., Boukhanovsky, A., Lees, M. H., & Sloot, P. M. A. (2020). Large-scale forecasting of information spreading. *Journal of Big Data*, *7*, [72]. https://doi.org/10.1186/s40537-020-00350-5

Journal of Big Data

RESEARCH Open Access

Large-scale forecasting of information spreading

Oksana Severiukhina[1*], Sergey Kesarev[1], Klavdiya Bochenina[1*], Alexander Boukhanovsky[1], Michael H. Lees[1,2,3] and Peter M. A. Sloot[1,2,3]

*Correspondence: oseveryukhina@gmail.com; k.bochenina@gmail.com
[1] ITMO University, St. Petersburg, Russia
Full list of author information is available at the end of the article

Abstract

This research proposes a system based on a combination of various components for parallel modelling and forecasting the processes in networks with data assimilation from the real network. The main novelty of this work consists of the assimilation of data for forecasting the processes in social networks which allows improving the quality of the forecast. The social network VK was considered as a source of information for determining types of entities and the parameters of the model. The main component is the model based on a combination of internal sub-models for more realistic reproduction of processes on micro (for single information message) and meso (for series of messages) levels. Moreover, the results of the forecast must not lose their relevance during the calculations. In order to get the result of the forecast for networks with millions of nodes in reasonable time, the process of simulation has been parallelized. The accuracy of the forecast is estimated by MAPE, MAE metrics for micro-scale, the Kolmogorov–Smirnov criterion for aggregated dynamics. The quality in the operational regime is also estimated by the number of batches with assimilated data to achieve the required accuracy and the ratio of calculation time in the frames of the forecasting period. In addition, the results include experimental studies of functional characteristics, scalability, as well as the performance of the system.

Keywords: Agent-based modelling, Model of information spread, Parallel simulation, Forecasting model, Data-driven model

Introduction

The purpose of this research is to understand and predict how information spreads in online social communities. It allows determining in advance the dynamic and characteristics of the processes. As a result, it becomes possible to detect deviations in behaviour that can help to identify unusual or suspicious activities like irrelevant, urgent and promotional content, or community hacking. One of the ways to achieve that is to build a computational model of how information process spreads over a definite topology for a certain set of rules of interacting agents' behaviour.

Models describing the processes of information dissemination in social networks can be divided into explanatory and predictive ones. Most of the models describe or forecast the dynamics of reactions to a single online message. However, these messages may be

Springer Open

presented in different contexts (e.g. thematic communities of Online Social Networks (OSN), such as a community of football fans or classical music lovers) and for various audience. Therefore, the model needs to be tailored to a given context.

We distinguish between several levels of modelling and forecasting information processes in OSN (Fig. 1):

- micro-level: information message (IM);
- meso-level: community (sequence of information messages in one community);
- macro-level: information spread in the form of IM between a set of communities.

At the micro level, the cascade on individual post dynamics is usually studied. At the meso-level, one may explore such things as influence of publication time on the impact of various messages and preferences of different segments of the audience. At the macro level, one may observe exchange of information between communities.

In this research, we present and evaluate an agent-based forecasting system for information spreading on micro- and meso-levels in the large-scale OSN. This system reproduces the aggregated impact of information messages from the individual agents' actions. To perform initial identification of model parameters, the retrospective data from OSN has been collected and the groups of agents and types of messages for a given community have been defined. During the forecast in the operational regime, we track the actual status of information process using web crawlers and feed this data into the model running at the supercomputer to tune the forecast in a real-time manner. This study investigates: (i) the possibility of reproducing the observed dynamics from individual reactions, (ii) the quality of forecasts in terms of accuracy and earliness, (iii) the impact of data assimilation on the quality of forecasts, (iv) the scalability and performance of the prediction system. Experimental studies performed on the Lomonosov supercomputer [1] using data collected during 2018 for two massive sets of news and charity communities.

The rest of the paper is organized as follows. Relevant information and related work are presented in Section 2. Section 3 describes the model, architecture of our system

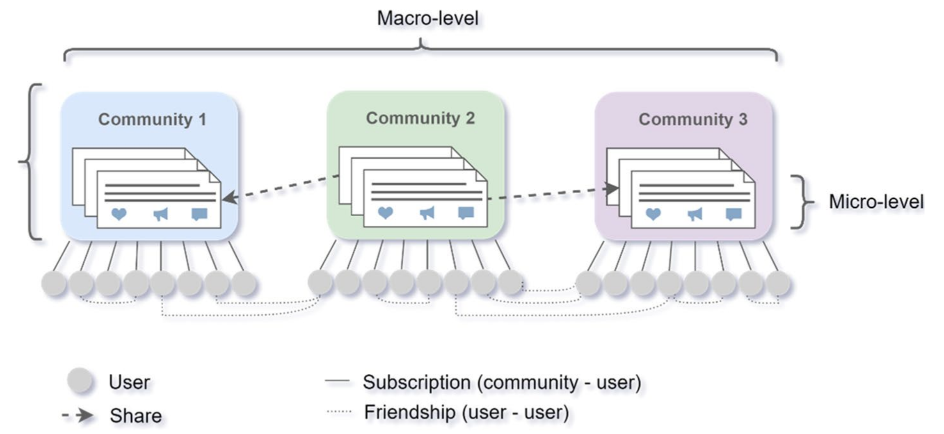

Fig. 1 Online social networks—micro, meso and macro levels

and methods for assessing forecast quality. Section 4 shows the details on the dataset, forecasting experiments, scalability and performance analysis. Section 5 presents conclusions and further research discussion.

Related research

To create data-driven models of process in a complex network, it is essential to understand two intertwined issues: (i) the most appropriate way to define agents (like individual users or a community), (ii) the structure and the parameters of the predictive models. In order to specify agents in the model, it becomes necessary to determine the features of their behavior, as well as the relationships between them.

The reaction of each user in a social network is unique and can be determined by their age, social status, preferences, the internal state of the user, the history of their interaction with the information source, and other characteristics. As users interact with various items over time, users' and items' features may change their behavior over time [2]. Moreover, the influence of media and social connections on the dynamics of user opinions should be noted [3]. Furthermore, the existence of the echo cameras effect in the network: polarized opinion in the nodes cluster leads to the diffusion of complex contagions, for example, fake news [4]. Some individual characteristics may influence processes in networks: heterogeneity of stateful agents [5] or agents' curiosity [6]. In works [7, 8] authors proposed an ontology-based approach to extract semantics of textual data and defined the domain of data. More precisely they semantically analysed the social data at the entity and the domain level. Proposed approach was evaluated with a public dataset collected from Twitter.

Although users differ in the number and ratio of responses to messages, OSNs may provide limited information about the activities of a single user (as opposed to properties of sub-populations of users). Therefore, creation of reaction models at the level of individuals is hampered both by the lack of data and the difficulty to distinguish and define all the factors that determine the users' response. The problem of reproducing agents' reactions can be solved by clustering users by their level of involvement, roles in the community, and parameters of their profile. For example, in [9] authors classify users into four groups: celebrities, organizations/media accounts, grassroots stars, and ordinary individuals. In our study, data-driven agent-based models are developed. Then, in the frames of these models, parameters of clusters from the history of individualized responses within a community are learned.

The topology of network affects on the information processes. Agent-based modelling approach allows investigating the bottom-up behavior or what-if analysis. In this case, agents (micro level) create emerging network behavior (macro level) [10]. In article [10], the authors propose an approach of complex agent networks that combines agent-based model and network approaches. Moreover, the social networks have the following "small world" effect, scale-free degree distributions and the modular structure [11]. The classical generative models like Erdos–Renyi, Watt-Strogatz and Barabasi-Albert models cannot reproduce all properties of real-world social networks. Moreover, the conductivity of links between entities may influence information dissemination: the users who have more common friends may have a greater possibility for the dissemination of information [12]. In our work, we use the real networks as inputs for the model.

Models of information spreading in networks can be divided into two categories: explanatory and predictive. In the explanatory models, information spreading is often considered in the same way as an epidemic spread process [13], where nodes can have one of the possible states in a concrete moment, for example, susceptible, infected, removed (SIR model). However, Weng [14] states that diseases differ from information: diseases spread as simple contagions, information spread as complex contagions, because the last one is affected by social reinforcement and homophily.

According to [13], there are three types of predictive models for OSN: the independent cascade model (ICM), the linear threshold model (LTM) and the game theory model. In the first type of model inactive node can be activated by the active node with some predefined probability. LTM model is a more complicated case, every interaction between nodes provides a cumulative effect on a node's state. In the last type of models, there are some specific restrictions and various agents' strategies. For example, the article [15] is aimed to explain how human factors impact on competitive information dissemination.

One of the sources of heterogeneity is the topicality of the message contents. Topic-aware independent cascade and topic-aware linear threshold models were proposed in work [16]. These model have different topics distribution and strength of nodes influence on each other depending on the topic. In other research, posts or news have a defined virality coefficient that depicts the popularity of the message and affects on the probability of sharing. In [14], the authors predict the virality of memes based on early spreading patterns in terms of community structure. In this case, investigation of activity patterns helps to detect viral memes [17].

In recent years, a significant number of papers are dedicated to more complex methods. Prediction of shares number based on temporal behavior patterns of users was studied in the article [18]. Machine learning approach with the passive-aggressive algorithm for predicting users' behavior in Twitter was proposed in [19]. Kefato at al. proposed a novel algorithm called CAS2VEC [20] that models information cascades as time series and discretizes them using time slices. In [21], the authors trained a probabilistic collaborative filter model to predict future retweets using Twitter data. In our work, we propose a forecasting method based on an agent-based model and data assimilation for increasing accuracy.

To forecast processes on large graphs, one needs to parallelize computation to be able to get the result of forecast in time. Parallelization can be applied to different steps of simulation from generative models (e.g. parallel Chung-Lu model [22]) to models of information spread (e.g. parallel SIR [23]). Moreover, different hierarchical synchronous parallel models for graph analytics can be used [24]. In this study, we modify our previous parallel algorithm for parallel simulation of dynamical processes on stochastic Kronecker graphs [25], to support arbitrary topologies and complicated models of agents' behavior.

Summarizing, our goal is to combine the advantages of complex agent network with data-driven approach learning the topology and parameters of agents from the data. To tackle the overall complexity of the resulting model and to adapt it to changing conditions, we constantly tune its parameters using data assimilation. Scalable parallel implementation is aimed to solve the problem of operational forecasting of information messages in OSN. This research is an attempt to propose whole methodology: from

Table 1 Forecast quality assessment' methods

#	Dataset	Features
1	followers	Community id, user id, date of collection
2	IMs	Community id, IM id, publication date, text, number of likes, shares and comments, date of collection
3	Likes:	Community id, IM id, user id, date of collection
4	Shares	Community id, IM id, reaction date
5	Comments	Community id, IM id, reaction date

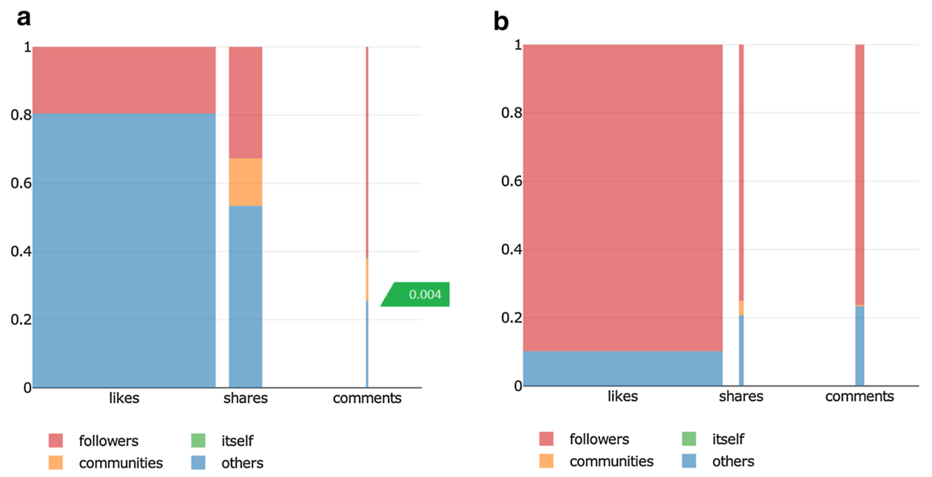

Fig. 2 Parts of different types of reactions for communities: **a** network charity, **b** news. The width of the rectangle shows the proportion of the reaction of this type, the height of the rectangle shows the proportion of reactions from various sources

retrospective data collection to getting the results of forecasts. In addition, proposed approach supports fine-tuning of the properties for various OSN contexts, temporal dynamics and different scales of simulation.

Dataset description

For this study, the data sets were collected from the VKontakte social network using the web crawler. The data includes charity community and news community (there are two types of IMs in this community, regular and IMs with advertisement) with IMs from April to May 2018 as historical data and several months from August as real-time data. The first community has 295 k followers and 100 IMs, second community—1900 k followers and 1500 k IMS.

The key features of the retrospective data are presented in Table 1. This dataset was used to train the basic parameters of the models.

Communities differ in both: the proportion of different reactions and the sources of reactions (Fig. 2). For instance, network charity community is distinguished by a high part of shares made by communities (these shares are done by administrators of individual charges or small communities), as well as a high part of reactions not from community subscribers, while in the news communities there is a larger activity of

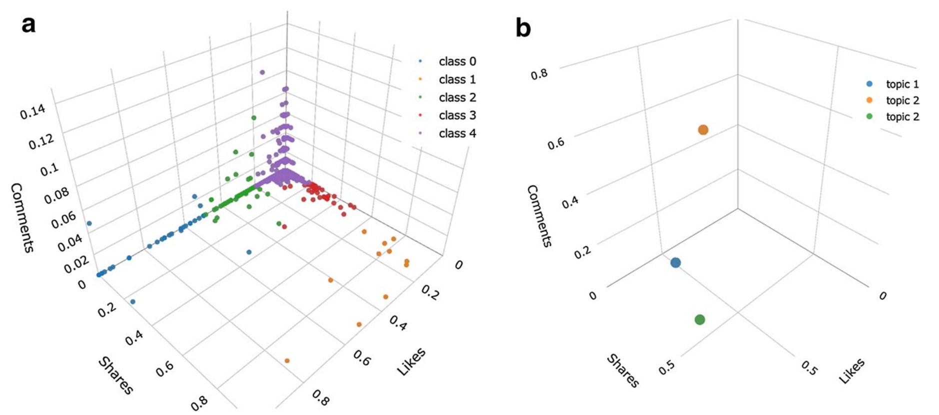

Fig. 3 Users' reaction types: **a** classes of users for a charity community, **b** user attitude to various topics in the reaction space

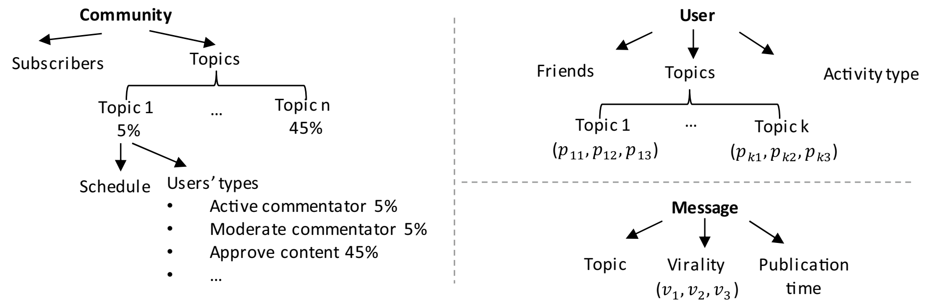

Fig. 4 Types of entities in the model and the attributes of entities

commentators. Figure 3b shows an example of a profile of user reactions containing the parameters of reactions to different topics. The selection of behavior' types was carried out in the space proportional to the selected types of reactions (Fig. 3a), e.g. point (0.2, 0.3, 0.6) means that the user left the response of the first type to 20% of IMs, the response of the second type to 30% of IMs and, finally, the response of the third type to 60% of the available IMs.

To sum up, as the output, we obtained the following characteristics: networks of communities and users and their actions to determine temporary activity and attitude to various topics.

The proposed method

Description of main entities in model

There are three main entities of online social networks: communities, users and IM (more detailed is in our previous study [23]). The network of entities in cyberspace is a directed graph, where the vertices of the graph are a set of communities and set of users, and edges are subscriptions or friendship links. Unit of information is IM, it can be transferred between vertices. Main characteristics of entities are presented in Fig. 4.

Every IM has a source of information and an identifier, time of publication, topic, as well as values of virality coefficient which differ for possible types of reaction on messages, for example, virality of sharing. We define virality as a potential impact of the message, which imply the probability of reaction (as an analogue for virality in epidemic spreading). A user is a vertex (receiver of information), which can respond to the received IM. For online social networks, three basic types of reactions are available: like (approval of content), comment (discussion of information), share (distribution of IM). Moreover, each user is described by types of daily/weekly activity, and a set of reactions to a set of IMs. A community is a network' vertex that broadcasts an IM to multiple users (community subscribers). It is described by a set of subscribers and a set of possible topics for IM. Each topic has the probability distribution of publication by the community depending on the time of day and day of the week.

The input parameters of the model are the network of subscriptions and friendships as well as parameters of internal models of communities and users. The network can be created artificially or extracted from the real social network. For the second case, we used the web crawler and collected information about subscribers of communities and friends of users. There are three internal models representing different drivers of information process and defining the behavior of entities: a model of IM's generation, a model of activity and model of reaction. The generative model of a community reproduces its publication activity. For a given community, it defines temporal patterns of publication (frequency during a day/a week) for IMs of different types according to the data presented on their page. The model of user activity determines the probability that the user will be online at different intervals of the day and may vary for different types of users (for example, early birds, night people). For this model, we used a number of user's reactions at different intervals of a day. The reaction model has parameters depending on the type of a user and IM, for different types of reactions. These parameters were set according to digital traces of user in different communities. All of the parameters above are estimated by using historical data from the social network. More details on the implementation of the internal models as well as setting of internal parameters may be found in [23].

Description of the general scheme

The architecture of the implemented system for operational forecasting of information processes in cyberspace is presented in Fig. 5. The web crawler allows efficiently collect data from various sources on the Internet, including the largest Russian online social network VKontakte. Depending on the chosen scenario, the crawler allows you to get two types of data: historical data for past time intervals and real-time data. Data collected by crawler are presented in JSON format with various parameters. However, further application of data in the model requires additional processing of the received files. Historical data are required to adjust the input parameters of the model. Data are stored as separate collections in MongoDB. The data obtained in this way can be processed and presented as input parameters for the model: network topology, communities' and users' parameters.

The forecasting model is responsible for two main tasks: (i) disseminating messages through the network, and (ii) refining the parameters when we get the data about the

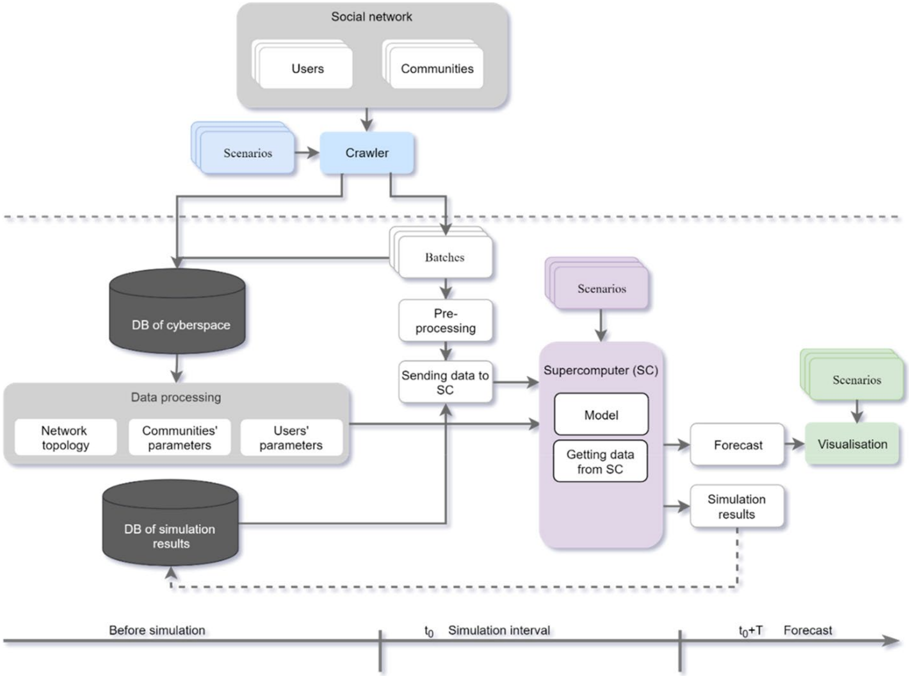

Fig. 5 General architecture of the system of operational forecasting of information processes in cyberspace

actual state of information process from the crawler. Parallel computations allow you to finish modelling on large-scale networks within a reasonable time. Reasonable means that we can obtain the forecast before it becomes outdated, and that we have enough time to react if this architecture is used for decision support system. The model was implemented on a C++language with MPI standard for message passing. The model uses the pattern of parallel communication Master/Slave. Master nodes are responsible for keeping statistics and IMs' generation. Slave nodes are responsible for hosting a sub-network and propagating IMs through it. A more detailed description of the algorithms and functionality of Master and Slave processes are presented in our previous article [26].

Real-time data are added to the model in the form of sequential batches that are processed and saved to a specific folder. Batch is a JSON file with the network data: each row or IM has the number of reactions of different types. In addition, it can include lists of users who actually reacted to the information. Every batch has an index starting from one.

During forecasting for one IM, the values of m predicted parameters for iterations 1, 2,..., n are calculated (from the current moment 0 to the forecast period T). A series of n values is recalculated every time a new batch of data from the crawling is received (Fig. 6). Thus, the result of a single prediction cycle is m· p · n · k values, where k is the number of batches. After applying new batch, internal parameters of the model can be specified for more accurate prediction by following options: resetting model time to batch time, changing the number of reactions, modification of users who viewed and reacted to the IMs, modification of the virality coefficients. A detailed description of the forecasting scheme was described in our previous work [27].

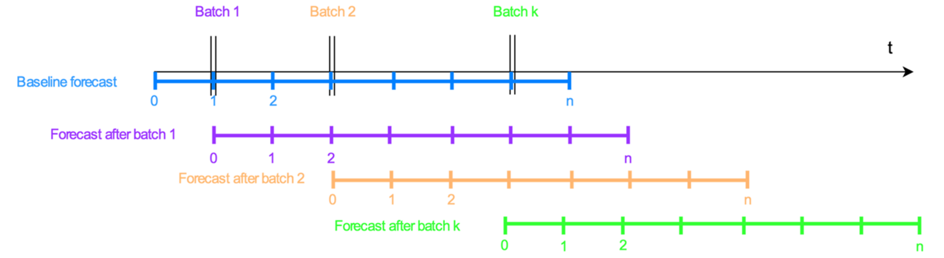

Fig. 6 Scheme of forecasting with data assimilation

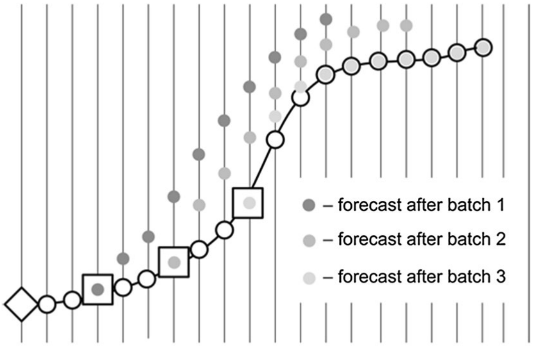
● – forecast after batch 1
● – forecast after batch 2
● – forecast after batch 3

Fig. 7 Scheme of the procedure for estimating the accuracy of the forecast with the assimilation of data (the rhombus indicates the moment of generation of IM, the squares are batches which are used both for assimilation and the evaluating the results (batches of assimilation), the circles are batches which are used only for evaluating the results (batches of evaluation), the circles are the response values for various forecast options)

Forecast quality assessment

The quality of the forecast is estimated in terms of accuracy, earliness and a number of batches needed to achieve desired accuracy of the forecast. Quality assessment is carried out according to the procedure illustrated in Fig. 7. After IM generation we can get information about a number of responses at different interval of time using batches. We use information from batches in two cases. Firstly, the evaluation of forecast obtained by our model (batches of evaluation). Secondly, assimilation of data from OSN and further recalculation of parameters in the model (batches of assimilation).

For forecasting at the micro level, the forecast accuracy for each IM is estimated by the MAPE, MAE metrics (Eqs. 1, 2) for an individual batch and is averaged over the set of batch files (Eq. 3).

$$MAPE(n) = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{1}$$

$$MAE(n) = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{2}$$

where A_t is actual value of the parameter at the iteration t, F_t is forecast value of the parameter at the iteration t.

$$MAPE(b,n) = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t^b - F_t^b}{A_t^b}\right| \tag{3}$$

where A_t^b is the actual value of the predicted parameter at the iteration t for the batch b, F_t^b. is the forecast value of the parameter at the iteration t for the batch b.

For forecasting at the meso-level, the accuracy of forecast for the series of IM is estimated with mean value (Eq. 4). Also, we measure the rate of improvement of forecast accuracy during data assimilation. It is estimated in terms of the number of batches needed toieve a given accuracy ε, at a fixed update rate ν (Eq. 5).

$$MAE(p,n) = \frac{1}{(p \cdot n)}\sum_{(i=1)}^{p}\sum_{(t=1)}^{n}|A_it - F_it|, \tag{4}$$

where A_{it} is the actual value of the predicted parameter at the iteration t. for IM i and F_{it} is the forecast value.

$$R(p,n) = \min(b) : \left(\frac{1}{p}\sum_{i=1}^{p}M(b,n)\right) < \varepsilon \tag{5}$$

where n. is number of iterations, p. number of IMs, b. is batch number, M. is selected metric to assess the accuracy of the forecast, ε is the minimal accuracy of the forecast.

To estimate the aggregated response to messages, e actual and model distributions of the number of reactions are compared using Kolmogorov–Smirnov criterion with significance level of 0.05 (Eq. 6).

$$\lambda = D_{N_1,N_2}\sqrt{\frac{N_1 + N_2}{N_1N_2}}, D_{N_1,N_2} = \sup_x \left|F_{1,N_1} - F_{2,N_2}\right| \tag{6}$$

where N_1, N_2 denote size of first and second samples respectively, $F_{1,N}(x)$ is empirical distribution function based on a sample size N.

Finally, earliness is estimated by the length of the time between the completion of the calculation and the end of the forecast period (Eq. 7), as well as the proportion of time allocated for the calculation to the length of the forecast period (Eq. 8).

$$z(b) = T - l \cdot b - \tau_b \tag{7}$$

$$z(b) = \frac{\tau_b}{T - l \cdot b} \tag{8}$$

where b is batch number, τ_b forecast time for current batch b, l is time between batch assimilation, T is forecast period.

The description of assessment methods is given in Table 2. The following notations for parameters are introduced: Δt is length of model time unit, n is number of iterations, T is prediction period ($n\Delta t = T$), p is number of posts, b is number of batches (A is set

Table 2 Forecast quality assessment' methods

#	What is evaluated	Forecast parameter	Method for assessme
1	IM response accuracy for the fixed virality value (without data assimilation)	Number of responses	Fix $\Delta t, n$. Calculate n. values of the predicted parameter (predicted parameter should be measured and be saved). Calculate $MAE(n)$, $MAPE(n)$. Calculate mean value of the metrics of accuracy er p IMs.
2	IM response accuracy with data assimilation	Number of responses	Fix T, n (in Fig. 7 $n = 9$). A series of forecasts for n. iterations is carried out: a) at the time of IM generation; b) after each batch from the set $b \in A$. batch of assimilation (with the adjustment of the model parameters from the obtained data). Figure 7 shows assimilation of three batches. Additionally, batches of evaluation $b \in C$. must be collected to ensure verification predicted values for all $\|A\| + 1$ forecasts. The averaged values of the forecast accuracy metrics for different assimilation batch are calculated.
3	Impact of the batches' frequency on the forecast accuracy	Number of responses	Calculate a series of averaged values of forecast accuracy metrics using the method (2) with a fixed length of the forecast period T, varying the length of the time interval l between batches
4	A number of batches required to achieve desired accuracy	Number of batches	Fix the desired accuracy ε, p, Δt, l, T, the maximum number of batches of evaluation $A_{\max}$. The assimilation of the batch is performed while the batch accuracy is less than ε or the maximum number of batches is exceeded $A_{\max}$. Accuracy assessment is made by the method (3)
5	Accuracy of the aggregated dynamics reproduction for the period	Number of responses	Fix, p, l, T. The response is predicted for p IMs for the period T. The numbers of responses in the model and the actual number of responses are measured. The value of λ is calculated by the Eq. (6). If$\lambda < \lambda'_{0.05} = 1.36$, then the null hypothesis of sample homogeneity is accepted
6	Earliness	Forecast time	Fix $\Delta t, l, T$, the number batches of assimilation k. For the moment of IMs generation ($t = 0$) and for each assimilation batch, the forecast is made for the period $T - l \cdot b$. The time of the forecast calculation $\tau_0, \ldots, \tau_k$, is measured. The series of values for the forecast earliness $z_0, \ldots, z_k$ is calculated using Eqs. (7, 8)

batches of assimilation, B is set batches of evaluation,$A \subseteq B$), l is the length of the time interval between batches (lk—the period of assimilation).

Results and discussion

Experimental study of the quality of forecasts

To test the functionality and the quality of forecasts, we consider two scenarios:

1. Prediction of the response to a single IM in the community (corresponds to the micro-scale modelling).

2. Prediction of the aggregated response to messages on various topics within a single community (corresponds to the meso-scale).

To study the quality of forecasts, experiments were carried out with parameters which are listed in Table 3. The results of the experiments were evaluated by the metrics described in Sect. 3.3. The quality of forecasting reactions to a single IM for the news community with 1,912,769 subscribers was investigated.

The accuracy assessments of the IMs response prediction for the basic value of virality and cases with the data assimilation are shown in Fig. 8 (methods 1—grey color and 2—blue color). For all types of reactions, the error decreases with an increase in the number of the batch (for likes, the median error less than 20% is reached on average in 1.5 h after the post publication, for shares and comments—after 3.5 h, due to the less number of these types of reactions). The error without assimilation (for basic virality averaged over historical data) is quite high (50–150%), due to the large variance of the actual viralities of the IMs. Moreover, Fig. 8 depicts the results of the assessment according to method 3 (of the impact of the frequency of assimilation batches). The median values of responses for different frequencies of receiving the batches are quite stable. The spread of the values for likes is smaller for more frequent batches; for rarer reactions (comments and shares), this trend is not observed.

To measure the accuracy according to method 4, we investigated the number of batches, which are needed to achieve desired accuracy (interval between batches is 30 min). Results for MAPE values from 0 to 1.5 are shown in Fig. 9a. To achieve an error value of less than 10% for likes and comments, about 10 batches are required, and 14 batches for shares, which are less predictable. Figure 9b shows dependences between forecast number and prediction period.

To study forecast quality for aggregated reactions on the set of IMs in communities, method 5 was used. Figure 10 provides a comparison of the aggregated dynamics of reactions for two types of IMs in the news community. Table 4 contains information on the results of calculating the statistics of the Kolmogorov–Smirnov criterion and the p-value for the two studied communities and three types of messages. For all the considered cases we cannot reject the null hypothesis that the samples have the same distribution.

The dynamics of forecasting time for a set of batches is shown in Fig. 11a. For the first batches, the time is growing, which relates to the need to update the states of the nodes of the complex network according to the batch data. However, starting from fourth batch, the time decreases due to the shorter forecast period. At the same time,

Table 3 Initial parameters for forecast quality assessment

Name	Description	Value
Δt	Length of model time unit	10 min
n	Number of iterations	72
T	Prediction period ($n\Delta t = T$)	720 min = 12 h (for each IM)
p	Number of IMs	30
k	Number of batches (A—set of assimilation batches, A—set of check batches,$A \subseteq C$)	$\|C\| = 24$, to study the influence of the batch frequency following number of batches for assimilation were used: for 30 min $\|A\| = 24$, for 60 min $\|A\| = 12$
l	The length of the time interval between batches (lk—the period of assimilation)	30 min ({30 min, 60 min, 90 min} to study the effect of the frequency of the batch)
N	The number of computational processes	8

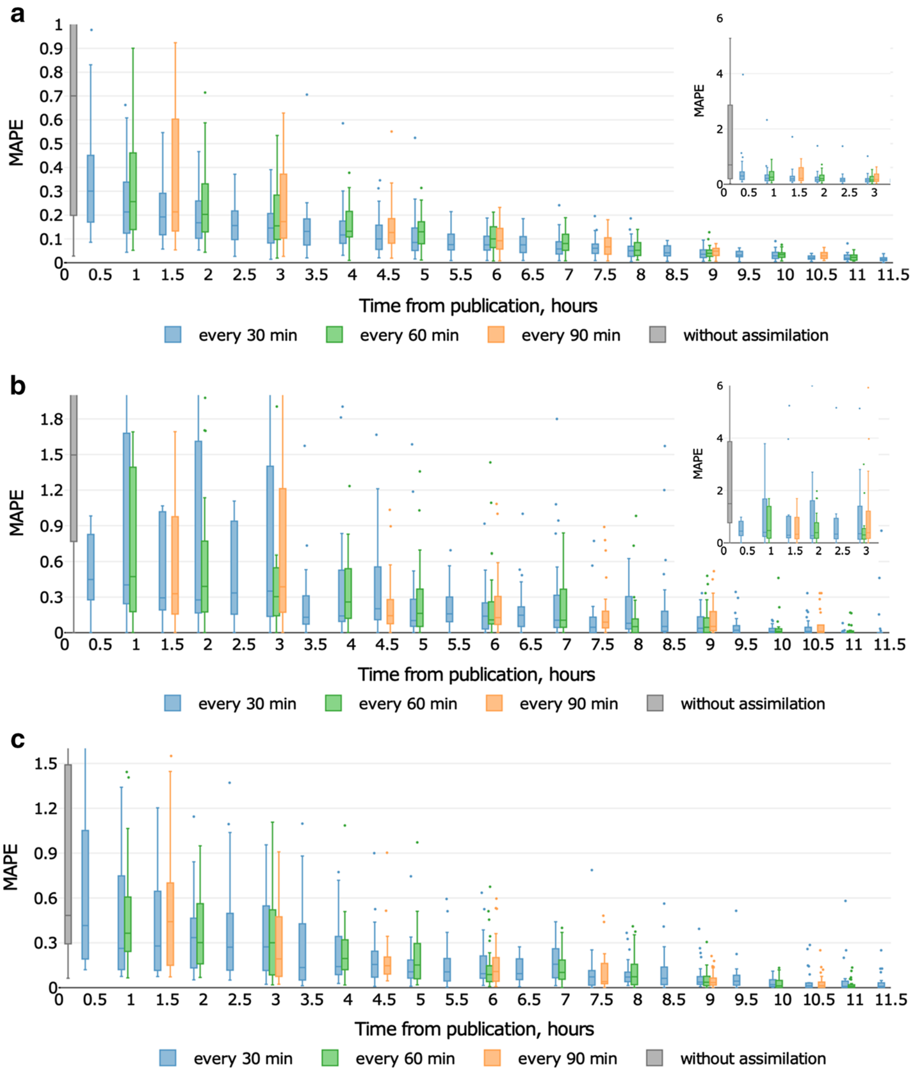

Fig. 8 The dynamics of the MAPE error (between actual and predicted value according to Eq. 3) depending on the batch number and the frequency batches of assimilation for different types of reactions: **a** likes, **b** shares, **c** comments. Zooms show a high value of errors for forecast without data assimilation

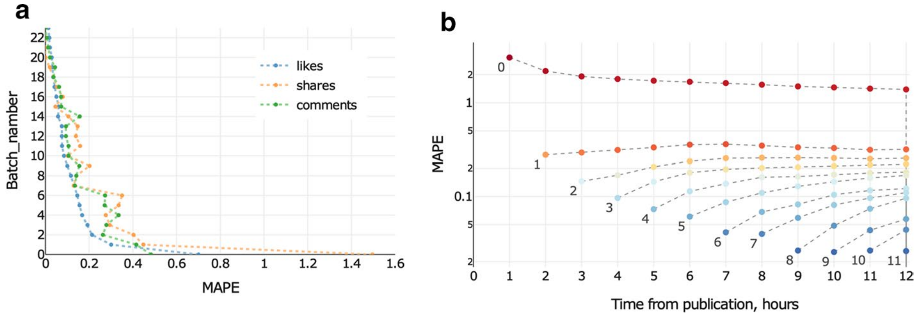

Fig. 9 Forecast accuracy **a** depending on the batch number, **b** depending on forecast hours (numbers of lines denote forecast number)

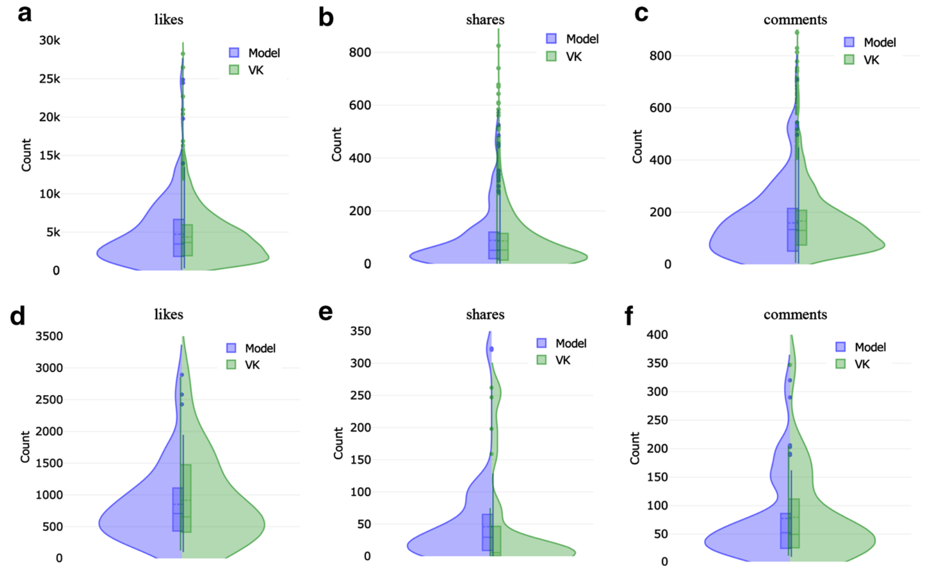

Fig. 10 Comparison of aggregated responses to messages for two types of IMs in news community: **a**, **b**, **c** for regular news; **d**, **e**, **f** for an advertising post

Table 4 Error metrics for the two types of thematic communities (statistic is the value of the statistics of the Kolmogorov–Smirnov criterion for two samples, critical value of statistic denoted maximum statistic value to determine that the samples have the same distribution)

Community type	Number of IMs	Critical value of statistic	Reaction type: like	Reaction type: share	Reaction type: comment
Charity community	$N_1 = 84$, $N_2 = 150$	0.185	Statistic = 0.14, p-value = 0.18	Statistic = 0.16, p-value = 0.09	Statistic = 0.07, p-value = 0.91
News community, regular IMs	$N_1 = 150, N_2 = 150$	0.157	Statistic = 0.14, p-value = 0.097	Statistic = 0.11, p-value = 0.27	Statistic = 0.12, p-value = 0.17
News community, ads	$N_1 = 24, N_2 = 50$	0.338	Statistic = 0.13, p-value = 0.91	Statistic = 0.32, p-value = 0.06	Statistic = 0.21, p-value = 0.39

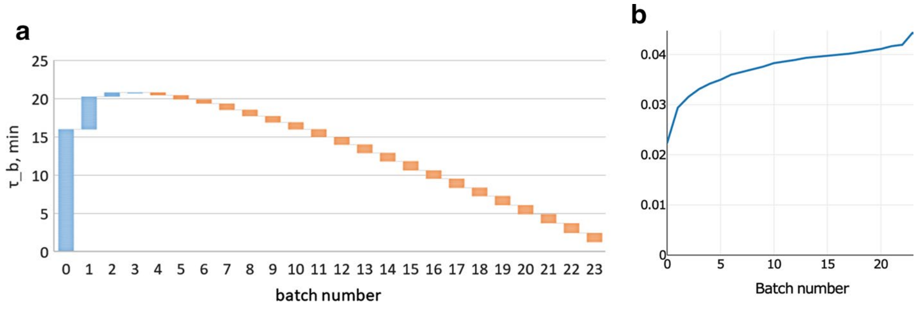

Fig. 11 Prediction time: **a** forecast calculation time, **b** earliness calculated by the Eq. 8

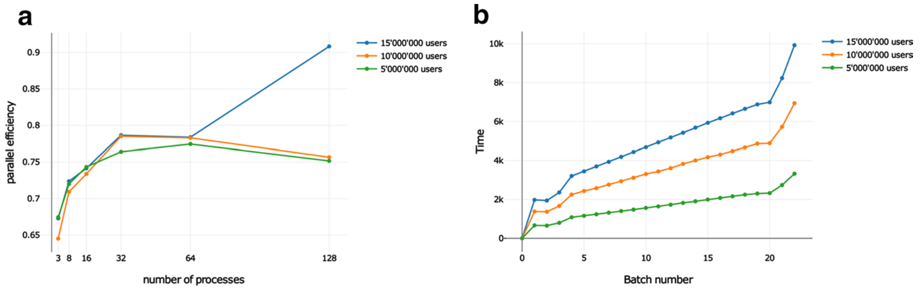

Fig. 12 Scalability results on Lomonosov supercomputer: **a** parallel efficiency, **b** simulation time depend on batch number

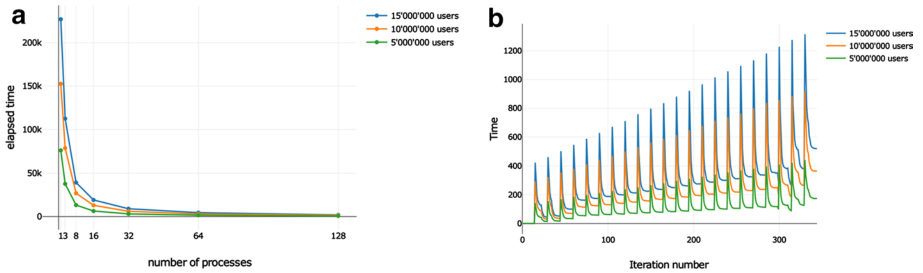

Fig. 13 Simulation time as a function of **a** number of processes, **b** iteration number (for 3 processes)

the earliness (Fig. 11b) are in the range of 0.02–0.04 (that is, the forecasting time ranges from 2 to 4% of the forecasting period), which shows the possibility of effective system operation in the operational mode.

Scalability and performance analysis

Figure 12 shows the results of the efficiency for the parallel implementation. The first experiment examines the scalability of predictive modelling in the parallel mode on the supercomputer Lomonosov. In this experiment, the size of one community's network is set to 5–15 M, and the number of IM is 5 or 10. In the data assimilation mode, 1 IM is added to each of 22 assimilation batches, and the number of processes varies from 1 to 128. Figure 12a demonstrates the parallel efficiency obtained in this experiment. For community sizes of 5 or 10 M nodes parallel efficiency varies from 0.64 to 0.78, peaking at 32 processes and then decreasing to 0.75 for 128 processes (Fig. 12a). For the graph with 15 M vertices, the parallel efficiency grows all the way up to 128 processes. The computational load (Fig. 12b) grows from batch to batch because every batch one new post is added to the system.

The second experiment estimates the performance of the predictive modeling module in the parallel mode on the supercomputer. Experiment parameters are the same as in the experiment above. Figure 13a shows the total simulation time for 22 batches of different sizes of MPI-communicator. Figure 13b shows the time taken by one iteration on 3 processes. In this experiment, the prediction period was set to 900 s, and the community size was 15 M. Running the module on three processes allows performing the forecast for 130% of the prediction period, 64 processes give 5% of the prediction period, and 128 processes give 2% of the prediction period. Therefore, the performance of the module is sufficient to obtain forecasts for large networks.

Conclusion and future works

We report on a novel multi-agent approach to predict the dissemination of information in online communities, taking into account both historical data and actual information about the states of messages in the social network at the current time. The data assimilation allows improve the quality of the forecast.

The forecast accuracy is estimated by MAPE and MAE for micro-scale, and the Kolmogorov–Smirnov criterion for aggregated dynamics (meso-scale). We also study the batches number (updates on current process state during simulation) needed to achieve the desired accuracy. Earliness is estimated by the time between the forecast calculation and the forecasting period.

The prognostic capabilities of the developed technology were assessed by the quality of forecasts using the developed methodology for two different thematic datasets (charity, news) and two different scales of cyberspace: micro-scale and meso-scale. Experiments were conducted in modes with and without the data assimilation. The median error for different types of user reactions reaches values less than 10% in 1.5–3 h after the message generation. The forecasted and empirical distributions of the reactions number to information messages (IM) in one context are indistinguishable by the Kolmogorov–Smirnov criterion for all experiments. The prediction time is 2–4% of the prediction period in a series of experiments. Performance studies show that the approach is scalable to very large networks.

The online monitoring mode of the system allows estimating real-world distributions some of the model parameters (such as IM virality, IM generation time, etc.). These distributions, estimated for specific contexts (like charity or news), allow predicting the information flow in advance and with high reliability. As a result it becomes possible to identify unusual or suspicious activities like the identification of fraudulent schemes based on the early detection of changes in audience engagement levels or activity patterns; identification of criminal networks based on the share of "suspicious" information sources in the general population of information sources; identification of suspicious stakeholder behavior based on the automatic estimation of expected virality and type of IMs.

Abbreviations

MAPE: The mean absolute percentage error; MAE: Mean absolute error; OSN: Online social networks; IM: Information message; ICM: Independent cascade model; LTM: Linear threshold model; SIR: Susceptible, infected, removed model.

Acknowledgements

The research was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

Authors' contributions

All mentioned authors contribute in the elaboration of the paper. All authors read and approved the final manuscript.

Funding

This research is financially supported by The Russian Science Foundation, Agreement №17-71-30029 with co-financing of Bank Saint Petersburg.

Availability of data and materials

Not available now.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] ITMO University, St. Petersburg, Russia. [2] Nanyang Technological University, Singapore, Singapore. [3] IAS, University of Amsterdam, Amsterdam, The Netherlands.

Received: 7 October 2019 Accepted: 27 August 2020
Published online: 03 September 2020

References

1. Sadovnichy V, Tikhonravov A, Voevodin V, Opanasenko V. "Lomonosov": Supercomputing at Moscow State University. Contemporary High Performance Computing: From Petascale toward Exascale (Chapman & Hall/CRC Computational Science). Boca Raton: CRC Press; 2013. p. 283–307.
2. Wang Y, Du N, Trivedi R, Song L. Coevolutionary latent feature processes for continuous-time user-item interactions. In: Advances in neural information processing systems; 2016. p. 4547–4555. http://papers.nips.cc/paper/6480-coevolutionary-latent-feature-processes-for-continuous-time-user-item-interactions
3. Quattrociocchi W, Caldarelli G, Scala A. Opinion dynamics on interacting networks: media competition and social influence. Sci Rep. 2014;4:1–7. https://doi.org/10.1038/srep04938.
4. Törnberg P. Echo Chambers and Viral Misinformation: modeling Fake News as Complex Contagion. PLoS ONE. 2017;13:1–23. https://doi.org/10.1371/journal.pone.0203958.
5. Zhu ZQ, Liu CJ, Wu JL, Xu J, Liu B. The Influence of Human Heterogeneity to Information Spreading. J Stat Phys. 2014;154:1569–77. https://doi.org/10.1007/s10955-014-0924-z.
6. Vega-Oliveros DA, Berton L, Vazquez F, Rodrigues FA. The impact of social curiosity on information spreading on networks. 2017. https://doi.org/10.1145/3110025.3110039
7. Wongthongtham P, Salih BA. Ontology-based approach for identifying the credibility domain in social big data. J Org Comput Electr Commerce. 2018;28(4):354–77. https://doi.org/10.1080/10919392.2018.1517481
8. Abu-Salih B, Wongthongtham P, Yan Kit C. Twitter mining for ontology-based domain discovery incorporating machine learning. J Knowl Manag. 2018;22:949–81. https://doi.org/10.1108/JKM-11-2016-0489.
9. Guo L, Wang W, Cheng S, Que X. Event-based user classification in weibo media. Sci World J. 2014. https://doi.org/10.1155/2014/479872.
10. Mei S, Zarrabi N, Lees M, Sloot PMA. Complex agent networks: an emerging approach for modeling complex systems. Appl Soft Comput J. 2015;37:311–21. https://doi.org/10.1016/j.asoc.2015.08.010.
11. Emilio F, Fiumara G. Topological features of online social networks. Commun Appl Ind Math. 2012. https://doi.org/10.1685/YYYYCAIMXXX.
12. Ou C, Jin X, Wang Y, Cheng X. Modelling heterogeneous information spreading abilities of social network ties. Simul Model Pract Theory. 2017;75:67–76. https://doi.org/10.1016/j.simpat.2017.03.007.
13. Li M, Wang X, Gao K, Zhang S. A survey on information diffusion in online social networks: models and methods. Information. 2017;8:118. https://doi.org/10.3390/info8040118.
14. Weng L, Menczer F, Ahn Y-Y. Virality prediction and community structure in social networks. Sci Rep. 2013;3:2522. https://doi.org/10.1038/srep02522.
15. Sun Q, Yao Z. Evolutionary game analysis of competitive information dissemination on social networks: an agent-based computational approach. Math Probl Eng. 2015;2015:1–12. https://doi.org/10.1155/2015/679726.
16. Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowl Inf Syst. 2013;37:555–84. https://doi.org/10.1007/s10115-013-0646-6.
17. Hui P-M, Weng L, Sahami Shirazi A, Ahn Y-Y, Menczer F. Scalable detection of viral memes from diffusion patterns. Cham: Springer; 2018.
18. Quan Y, Jia Y, Zhou B, Han W, Li S. Repost prediction incorporating time-sensitive mutual influence in social networks. J Comput Sci. 2018;28:217–27.
19. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. Proc. Fifth Int. Conf. Weblogs Soc. Media - ICWSM'11. 586–589 (2011).
20. Kefato, Z.T., Sheikh, N., Bahri, L., Soliman, A., Montresor, A., Girdzijauskas, S.: CAS2VEC: Network-Agnostic Cascade Prediction in Online Social Networks. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 72–79. IEEE (2018).
21. Zaman TR, Herbrich R, Van Gael J, Stern D. Predicting Information Spreading in Twitter. Proc Comput Soc Sci Wisdom Crowds Work. 2010;55:1–4.
22. Alam M, Khan M. Parallel algorithms for generating random networks with given degree sequences. Int J Parallel Program. 2017;45:109–27. https://doi.org/10.1007/s10766-015-0389-y.
23. Bhatele A, Yeom JS, Jain N, Kuhlman CJ, Livnat Y, Bisset, KR, Kale LV, Marathe MV: Massively parallel simulations of spread of infectious diseases over realistic social networks. In: Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017. p. 689–694 (2017).
24. Liu S, Chen L, Li B, Carnegie A. A Hierarchical Synchronous Parallel Model for Wide-Area Graph Analytics. In: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications. pp. 531–539. IEEE (2018).
25. Bochenina K, Kesarev S, Boukhanovsky A. Scalable parallel simulation of dynamical processes on large stochastic Kronecker graphs. Futur. Gener. Comput. Syst. 2017;78:502–15. https://doi.org/10.1016/j.future.2017.07.021.
26. Kesarev S, Severiukhina O, Bochenina K. Parallel simulation of community-wide information spreading in online social networks. Sci: Commun Comput Inf; 2018.
27. Severiukhina O, Kesarev S, Petrov M, Bochenina K. Parallel forecasting of community-wide information spread with assimilation of social network data. Procedia Comput Sci. 2018;136:228–35. https://doi.org/10.1016/j.procs.2018.08.260.

Publisher's Note