CrossSimON: A Novel Probabilistic Approach to Cross-Platform Online Social Network Simulation

Jinwei Liu[1], Wingyan Chung[1,2,3], Yifan Huang[1], and Cagri Toraman[1]

[1]School of Modeling, Simulation, and Training, University of Central Florida, Orlando, Florida 32826, U.S.A.
[2]Department of Computer Science, Faculty of Engineering, The University of Hong Kong, Pokfulam, Hong Kong
[3]Dept. of Decision Sciences & Managerial Econ., School of Business Admin., The Chinese University of Hong Kong, Shatin, Hong Kong
Email: jliu@ist.ucf.edu, wchung@ucf.edu, yifan@knights.ucf.edu, cagri.toraman@ucf.edu

***Abstract*—The increasing popularity and diversity of online social networks (OSNs) have attracted more and more people to participate in multiple OSNs. Learning users' behavior and information diffusion across platforms is critical for cyber threat detection, but it is still a challenge due to the surge of users participating in multiple social platforms. Existing research on profile matching requires user identity information to be available, which may not be realistic. Little prior research payed attention to mapping behavioral patterns across platforms. We designed and implemented an efficient two-level probabilistic approach called CrossSimON to mapping user-group behavior across platforms. CrossSimON considers the activity level and network position at both individual user level and group level to correlate activities across social platforms. To evaluate the effectiveness of CrossSimON in modeling social activity across platforms, we conducted experiments on three online social platforms: GitHub, Reddit and Twitter. Our experimental results show that CrossSimON outperformed the Benchmark in 3 out of 5 simulation metrics. CrossSimON achieved better performance in user activity prediction. The research provides new strategy for cross-platform online social network simulation, and new findings on simulating OSNs and predictive analytics for understanding online social network behavior.**

Keywords*— Social media, cross-platform simulation, group mapping, similarity, Reddit, regression, social networks, information spread.

I. Introduction

With the proliferation of online social media and social network, users have been introduced to many online social platforms such as Twitter, GitHub, Reddit, or Instagram. According to the 2018 survey of U.S. adults conducted by Pew Research Center, 73% of Americans use two or more of the eight social media platforms (Twitter, Instagram, Facebook, Snapchat, YouTube, WhatsApp, Pinterest and LinkedIn) [1]. The prevalent use of social media facilitates the spreading of malicious/criminal information easily across platforms. Learning users' behavior and information diffusion across platforms is important for cyber threat detection, such as tracing cybercriminals [2], [3].

Information and rumors can travel across different online social media platforms, and linking users or entity identities across platforms can help better understand information diffusion [4]–[6]. However, user identity information for the same person in real world can be very different across different online social platforms [7], [8]. Also, online social media data is typically huge, noisy, incomplete and highly unstructured [9].

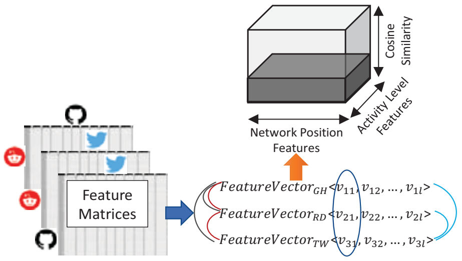
Cosine Similarity
Activity Level Features
Network Position Features
Feature Matrices
$FeatureVector_{GH} < v_{11}, v_{12}, \ldots, v_{1l} >$
$FeatureVector_{RD} < v_{21}, v_{22}, \ldots, v_{2l} >$
$FeatureVector_{TW} < v_{31}, v_{32}, \ldots, v_{3l} >$

Fig. 1. Multi-level features for group identity and user identity.

Existing literature neglects the mapping of behavioral patterns of both users and the groups they belong to across platforms. Also, most of the prior approaches fall into one of the following categories: profile attribute-based approach, network structure-based approach, content-based approach, and timestamp-based approach. However, the solutions in the literature have some limitations (e.g., Network-based user identification poses several major challenges such as unavailable entire network structure due to the large scale and privacy issues maintained by online social media companies [10]–[12]) and cannot well simulate the online social behavior across platforms.

Literature has shown that group behavior correlates with individual behavior [13]–[15]. Inspired by this, we consider both user identities and group identities and propose CrossSimON to mapping user-group behavior across platforms. CrossSimON considers the activity level and network position at both individual user level and group level (see Figure 1) to correlate activities across social platforms. CrossSimON provides a fine-grained solution to simulate online social behavior across platforms. Our experimental results show that CrossSimON can better improve the performance of cross-platform social behavior simulation.

II. Related Work

Information spread across platforms (e.g., OSNs) has become an active research area. This line of research includes cross-platform user identification and cross-platform topic modeling.

A. Cross-Platform User Identification

Many studies investigate user identification across platforms. Jain *et al.* [16] proposed two user identity search algorithms based on profile attributes of a user, and they used

978-1-7281-2504-6/19/$31.00 ©2019 IEEE

the two search algorithms to find user identity on Facebook, given her identity on Twitter. Roedler *et al.* [17] proposed a content driven profile matching approach to map users across different OSNs. Zhou *et al.* [18] proposed the Friend Relationship-Based User identification (FRUI) algorithm for addressing the problem of user identification across OSNs. FRUI calculates a match degree for all candidate User Matched Pairs (UMPs), and only UMPs with top ranks are considered as identical users. Goga *et al.* [19] proposed a set of techniques for user identification (identifying accounts on different social network sites that all belong to the same user) across online social network sites based on innocuous information (e.g., location and timing patterns). Zafarani *et al.* [20] proposed a method to match profiles across multiple social network sites just by analyzing user names. The approach can work only if a user chooses deliberately to use the same user name. Zhang *et al.* [21] presented a two-phase network alignment framework UMA (Unsupervised Multi-network Alignment) to identify the connection between shared users' accounts in multiple social networks (i.e., the anchor links).

B. Cross-Platform Topic Modeling

Some studies investigate applying topic models on multiple social media platforms. Guo *et al.* [22] proposed a model that considers social-relationship among users for topic modeling and applied the model on SinaWeibo and Twitter datasets. Cho *et al.* [23] designed a model that incorporates users' social interactions and attributes for topic modeling and applied the model on six social media platforms [5]. However, many of these works perform the topic analysis on each platform independently without linking users across platforms. There are also some works that study cross-platform topic modeling. Liu *et al.* [24] introduced TopicPanorama that focuses on visualisation of topics across multiple social media sites. Lee *et al.* [25] focused on linking user-generated contents across multiple social media sites to understand the cross-platform content posting behavior of a particular user. Pokharel *et al.* [26] proposed a framework referred to as Multi-source Social Topic Media Analysis (xSMA) framework to model, rank and semantically analyze emerging topics across various social media platforms.

However, the above works either focus on user identification (e.g., user profile mapping) or content mapping across platforms. Also, most of the above works can only generate the mapping between users & contents. None of existing works consider the social networks of the online communities.

III. CrossSimON: An Approach to Cross-Platform Simulation of OSN Activities

In this paper, we consider both user identities and group identities and propose CrossSimON (Cross-platform Simulation of Online Social Networks) to mapping user-group behavior across platforms. Unlike prior works, CrossSimON considers the activity level and network position at both individual user level and group level to correlate activities across social platforms. CrossSimON enables both group linkage and user linkage. CrossSimON first utilizes the multi-level features for group linkage and then associates the users with groups based on a novel probabilistic approach. In the following, we first introduce the "group" concept and the rationale of our model, and then we formulate our problem. Finally, we present the design of CrossSimON.

TABLE I
Notations.

l	# of features in feature vector	P_i	Platform i
g_{ij}	The jth group in P_i	N_{P_i}	# of event types in P_i
u_{ik}	The kth user in P_i	t	Time t
$p(u_{ik}\|g_{ij},t)$	Prob. of u_{ik} being associated with g_{ij} given g_{ij} and time t	U_{ij}	A set of users who participate in g_{ij} at t
E_{jr}	Counts of event type r in P_j	$\beta_{r,m}$	Regression coefficient
$\hat{E}_{im}$	Predicted count of event type m in P_i	A_{ij}	Action type j in P_i

A. Problem Statement

We assume that the individual user behavior correlate with group behavior. To simulate the information across social platforms, we define *group* as a (possibly evolving) information artifact which is read, forwarded, commented, edited or otherwise manipulated by the users. The rationale of introducing group for cross-platform online social behavior simulation results from one of the most striking phenomena in the natural world: collective behavior of social animals, particularly coordinated group movements [13]–[15]. Table I shows the main notations used in the paper.

Problem Statement: Given the training data from three social platforms (one target platform and two source platforms) and the testing data (concurrent data which are comprised of the data from the other two source platforms) from two source platforms, how to simulate the activity (i.e., sequence of events including the information of nodeID, nodeUserID, actionType, nodeTime, groupID, parentID, rootID) of the target platform in the testing period (the period of the concurrent data)? How to simulate the activity of the target platform in the testing period with a gap between the training data and testing data?

B. The Design of CrossSimON

Figure 2 shows the framework of the CrossSimON. CrossSimON consists of three major components: (1) group mapping based on multi-level features (activity level and network position), (2) user identification using the probabilistic approach, (3) cross-platform simulation using multiple linear regression. Below we present the design of each component.

1) Group Mapping: In training period, we first perform feature alignment across platforms because different social platforms have different structures and strategies for presenting user/entity profiles (e.g., node types and link types). Then, we use the extracted group features to find the similar group in each source platform given the group in the target platform based on the cosine similarity in Formula (1) and the timestamp of the groups.

$$sim(g_{i,q}, g_{j,v}) = \frac{S \cdot T}{||S||\ ||T||} \tag{1}$$

where S and T are the feature vectors of group $g_{i,q}$ and $g_{j,v}$, respectively. Figure 3 shows the design of group mapping, which maps the groups across platforms based on the cosine similarity derived from the group features. CrossSimON chooses the group that has the highest cosine similarity score among the candidate group list as the similar group.

• Feature Selection: We extract both activity level features and network position features for the entire training period. In each platform, we extract the groupID, timestamp, the group features such as ratio of the number of the event type A_{ij} (an action/event type such as CreateEvent in GitHub, Post in Reddit, Tweet in Twitter, etc.) to the maximum number of event type A_{ij} from the group, degree centrality of the group, closeness centrality of the group, and eigenvector centrality of the group. The activity level features for group include: the ratio of the number of the event type A_{ij} (such as CreateEvent in GitHub, Post in Reddit, Tweet in Twitter, etc.) to the maximum number of the event type A_{ij} from the group, and the network position features include: degree centrality of the group, closeness centrality of the group, eigenvector centrality of the group. To compute the dot product of the feature vectors of two groups from two platforms, we perform feature alignment across different platforms. Specifically, we align the feature “numCreateToMaxEvts” (ratio of the number of CreateEvents to the maximum number of CreateEvents from the group) with “numSubToMaxSub” (ratio of the number of Posts to the maximum number of Posts from the group) in Reddit and “numTweetsToMaxTweets” (ratio of the number of Tweets to the maximum number of Tweets from the group) in Twitter; we align the average ratio of the number of the other 9 event types (see Section IV-A for the explanation of event types) from the group to the maximum number of the other 9 events from the group in GitHub with that of the Comment event in Reddit and that of the other 3 event types in Twitter; we align the degree centrality of the group, closeness centrality of the group, eigenvector centrality of the group in GitHub to that of Reddit and that of Twitter, respectively.

2) User Identification Using the Probabilistic Approach: After finding the mapping of groups across different platforms, CrossSimON then identifies the user associated with the group (i.e., the user who participates in the group) based on the probabilistic approach. Below we use cryptocurrency domain as an example to introduce the details of the approach.

Denote g_{ij} $(i = 1, 2, 3)$ as the jth group in platform P_i. Denote u_{ik} as the kth user in platform P_i. Let $p(u_{ik}|g_{ij}, t)$ be the probability that the user associated with the group g_{ij} is u_{ik} given that the group is g_{ij} and the time is t, and the probability $p(u_{ik}|g_{ij}, t)$ can be represented as the number of occurrences that the user associated with the group g_{ij} is u_{ik} given that the group is g_{ij} and the time is t (denoted by $n_{ikj,t}$) divided by the total number of occurrences that the user associated with the group g_{ij} is u_{ih} given that the group is g_{ij} and the time is t (denoted by $n_{ihj,t}$), where u_{ih} could be any user in the training period. Hence, we have

$$p(u_{ik}|g_{ij}, t) = \frac{n_{ikj,t}}{n_{ihj,t}} \tag{2}$$

where u_{ih} represents a particular and arbitrary user that is associated with the group g_{ij} at time t. The higher the value of $p(u_{ik}|g_{ij}, t)$, the higher the probability that the user u_{ik} participates in group g_{ij} at time t. Hence, we consider the user with the maximum $p(u_{ik}|g_{ij}, t)$ as the user (the user who participates in the group g_{ij} at time t) associated with the group g_{ij} at time t. Hence, we have

$$u_{ik} = \underset{u_{ik} \in U_{ij}}{\arg\max}\ p(u_{ik}|g_{ij}, t) \tag{3}$$

where U_{ij} is a set of users consisting of all the users who participate in the group g_{ij} at time t. Therefore, Formula (2) and Formula (3) can be used to identify the user associated with the group g_{ij} at time t.

3) Cross-Platform Simulation Using Multiple Linear Regression: We assume that there is no time lapse between two platforms' activities. After finding the users associated with the groups based on the probabilistic approach, CrossSimON uses the multiple linear regression model to predict the counts of different event types in the target platform given the concurrent data from the other two source platforms. After generating the predicted counts of different event types, CrossSimON generates the predicted sequence of event types and the simulation of the cross-platform online social networks.

To predict the counts of events in platform i (P_i), we build the following multiple linear regression model shown in Formula (4)

$$\hat{E}_{im} = \beta_{0,m} + \sum_{r=1}^{N_{P_j}} (\beta_{r,m} \cdot E_{jr}) + \sum_{s=1}^{N_{P_k}} (\beta_{N_{P_j}+s,m} \cdot E_{ks}) \tag{4}$$

where $\hat{E}_{im}$ is the predicted count of event type m in P_i (platform i), E_{jr} represents the count of event type r in P_j, E_{ks} represents the count of event type s in P_k, N_{P_j} and N_{P_k} are the number of event types in P_j and P_k, respectively, $\beta_{u,m}$ $(u \in \{0, 1, ..., N_{P_j} + N_{P_k}\})$ is the regression coefficient

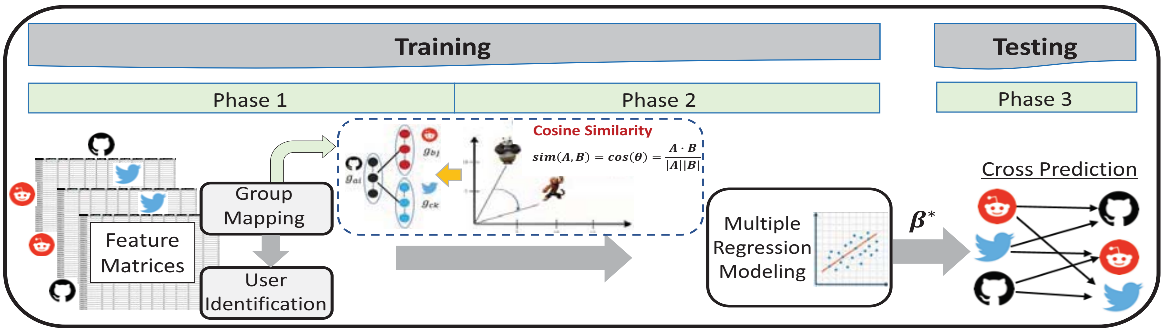

Fig. 2. Framework of the CrossSimON for cross-platform online social network simulation.

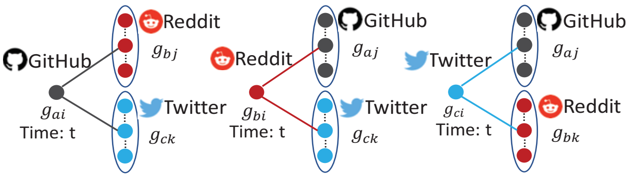

Fig. 3. Group mapping across different social media platforms.

of the multiple linear regression model. Each time we use the regression coefficients with the same indexed hours in one week, i.e., $0, ..., 167$ (0 and 167 represent the first hour and the last hour of the week, respectively) as that of the group and user, and each event type for making prediction.

Algorithm 1 shows the pseudocode for CrossSimON algorithm. The algorithm first aligns features (including activity level features and network position features) across platforms and computes the cosine similarity between the groups from the source platform and target platform (line 1), and maps the groups across platforms based on the cosine similarity score and the group's timestamp (line 2). Next, the algorithm computes the probability of a user (e.g., u_{ik}) being associated with a group g_{ij} given that the group is g_{ij} and the time is t (line 3), then the algorithm identifies the user who belongs to the group based on Formulas (2) and (3) (line 4). Next, the algorithm uses the scikit-learn package to generate the regression coefficients (line 5). After getting the regression coefficients, the algorithm uses the multiple linear regression to predict the counts of event types of the testing period in the target platform (line 6). Finally, the algorithm generates the simulated sequence of events of the testing period in the target platform based on the predicted counts of event types (line 7). This algorithm can be applied to generate the simulation for all the platforms (GitHub, Reddit and Twitter) and all the domains (cryptocurrency, software vulnerability, cyber threat). Due to the page limit, we choose cryptocurrency domain and choose Reddit as the target platform, and present our experimental results in Section V.

Algorithm 1: Pseudocode for CrossSimON

```
Input: Training data (including the extracted activity level features and
       network position features, regression coefficients, hourly event
       counts, information of nodeID, nodeUserID, action Type, etc.)
       from multiple ($m$) platforms and concurrent data (testing data)
       from two source platforms
Output: The simulated sequence of events of the testing period in the
       target platform
1 Align features across platforms and compute the cosine similarity
2 Map groups across platforms based on the cosine similarity and
    timestamp of the groups    //Section III-B1
3 Compute the probability of a user $u_{ik}$ being associated with a group
    $g_{ij}$ given that the group is $g_{ij}$ and the time is $t$   //Formula (2)
4 Associate users with the groups based on the calculated
    probabilities   //Formulas (2) and (3)
5 Generate the regression coefficients using the scikit-learn
6 Predict the counts of event types of the testing period in the target
    platform using multiple linear regression   //Formula (4)
7 Generate the simulated sequence of events of the testing period in the
    target platform based on the predict counts of event types
```

IV. Applications

In this paper, we build CrossSimON based on three popular online social networks: GitHub, Reddit and Twitter. Below, we first introduce these three social platforms, then we describe the dataset considered in CrossSimON.

A. Social Platforms in CrossSimON

The GitHub social network consists of two types of nodes: users and software repositories, and ten types of links: events between a user and a repository. These ten types are "Watch", "Fork", "Pull", "Push", "Issue", "Create", "Delete", "PullRequest", "IssueComment", "PullRequestReviewComment" and "CommitComment". A "Watch" event occurs when a user clicks the "Watch" button on GitHub to watch a repository, etc. The Reddit social network consists of two types of nodes: users and subreddit, and two types of links: events between a user and a subreddit. The two types are "Post" and "Comment". A "Post" event occurs when a user makes a submission (e.g., sharing a link, etc.). The Twitter social network consists of one node type users and four types of links (events): "Quoted Tweet", "Retweet", "Tweet" and "Reply". "Quoted Tweet" enables a user to say something along with his/her Retweet, while showing people the original tweet.

B. Dataset Description

The data considered for the development of CrossSimON, consists of eight-month training data and around half month testing data. The raw data consists of hourly activities on GitHub, Reddit and Twitter. In the data preprocessing phase, necessary information in the form of user-id, group-id (e.g., repository-id in GitHub, Subreddit-id in Reddit, Hashtag in Twitter) and event-type with timestamps were extracted from these hourly activities and aggregated per day. To enable the cross-platform prediction and simulation, we process the concurrent data which are comprised of the data from the other two source platforms (e.g., The concurrent data for generating the predicted results in GitHub are the testing data from Reddit and Twitter, and the concurrent data for generating the predicted results in Reddit are the testing data from GitHub and Twitter; the concurrent data for generating the predicted results in Twitter are the testing data from GitHub and Reddit). The period of the concurrent data for GitHub is December 1st to 14th 2017, the period of the concurrent data for Reddit is December 22nd to 28th 2017, and the period of the concurrent data for Twitter is January 1st to 15th 2018.

V. Experiments and Findings

In this section, we first describe the metrics and measurements used for evaluation, then present the experiment setup, and finally present our experimental results, findings and analyses.

A. Performance Measurements and Metrics

To evaluate the performance of CrossSimON, we primarily focus on three evaluation metrics: Absolute Percentage Error (APE), Jensen-Shannon Divergence (JSD), Normalized Root Mean Square Error (NRMSE). APE measures the size of the error in percentage terms. JSD measures the similarity between two probability distributions. NRMSE is a fraction of the overall range that is typically resolved by the model, e.g., $\frac{RMSE}{Q_1 - Q_3}$, where RMSE (Root Mean Square Error) is the measure of the differences between values (sample or population values) predicted by a model and the values observed. The lower value of these metrics indicates better performance.

To test the performance of CrossSimON, we selected the following representative measurements: comm_modularity (community modularity which measures how strongly the network resolves into communities or modules), comm_num_user_actions (average number of daily user actions for users within the community), user_activity_dist (user activity distribution which measures the distribution over user activity for all users), user_unique_content (measuring number of unique events that users interact with, e.g., number of unique Posts that users comment on), cont_activity_disp_geni_comm (the disparity of total level of activity among users), that can be categorized into three categories: community, user and content. Community measurements evaluate the quality of information spread over communities (i.e., represented as subreddits). User measurements evaluate the quality of how information is spread among users. Content measurement evaluates the quality of how to determine if some users have more influence over how information spreads.

B. Experiment Setup

1) Data Collection: Our dataset consists of Github, Reddit, and Twitter activities in the cryptocurrency domain. Training dataset includes 557,643 Github activities recorded in 9,115 repositories; 1,321,279 Reddit activities recorded in 9 subreddits; 7,381,980 Twitter activities recorded in 1,405 hashtag groups. The training data were related to activities of 65,628 Github, 80,575 Reddit, and 516,117 Twitter users who participated during the time period from January 1 to August 31, 2017. Simulation dataset includes 142,755 Reddit activities recorded in 3 subreddits. The simulation data were related to activities of 29,565 Reddit users who participated during the time period from December 22 to December 28, 2017.

- Feature Description: In Reddit platform (https://www.reddit.com/), a "subreddit" is a forum dedicated to a specific topic. A Post (also called "submission") is a top-level news item submitted by a user. Other users may respond to a Post by submitting Comments and by commenting on existing Comments. Posts and Comments then form a hierarchical structure within their respective subreddit. The Posts, Comments, and temporal interactions among users in subreddits form bi-partite undirected networks, in which two types of nodes are user and subreddit, and two types of links are Comment and Post. A Comment is a link between two users whereas a Post is a link between a user and a subreddit. In GitHub platform (https://github.com/), users can share and edit software repositories. A repository (repo for short) is an independent project where users store code, documentation, resource files, and references. The ten types of events, and temporal interactions among users in repos form bi-partite undirected networks, in which two types of nodes are user and repo, and ten types of links are then event types (e.g., "Watch", "Fork", etc.). In Twitter platform (https://twitter.com/), users can communicate through "tweets" or messages up to 140 characters long. The four event types, the interactions among users form a single network, in which one node type is user and four types of links are "Quoted Tweet", "Retweet", "Tweet" and "Reply".

2) Experiment Design: We compared CrossSimON with the Benchmark, which generates a simulation sequence by replicating the same users and events occurred prior to the simulation period. A simulation sequence consists of all events predicted by a method, and each event specifies nodeID (ID of Post/Comment), nodeUserID (ID of the user who posted the Post/Comment), actionType (event type), nodeTime (timestamp of when the Post/Comment was posted), groupID (ID of a subreddit), parentID (ID of the Post/Comment to which the specified nodeID is an immediate reply), rootID (ID of the original post). We used events occurred in the ground truth to compare against the simulated events to measure the performance of a method. In practical scenarios, a gap typically exists between the training period and testing period [27]. To test the effectiveness of CrossSimON in handling the gap between the training period and testing period, we chose the period from December 22 to December 28 as the testing period (when the target platform is Reddit), and used the data in the period from March 25 to March 31, 2017 and the period from August 25 to August 31, 2017 to generate the results for Benchmark approach, respectively (The end date of CrossSimON's training period is the same as that of Benchmark). The average values of experimental results of these two periods are reported in Table II.

C. Findings and Analyses

Table II shows the experimental results of simulating Reddit events with GitHub and Twitter activities over the period of December 22-28, 2017. CrossSimON outperforms the benchmark method on the metrics of user_activity_dist, user_unique_content, cont_activity_disp_geni_comm. Also, the performance of CrossSimON on other measurements (e.g, comm_num_user_actions) is close to that of the Benchmark: the difference between the APE value of comm_modularity for CrossSimON and that for Benchmark is only 7.963, and the difference between the JSD value of comm_num_user_actions for CrossSimON and that for Benchmark is only 0.343. This indicates that CrossSimON performs well on some of the measurements. The effects of the gap between the training period and testing period does not compromise the performance of CrossSimON much. This suggests that CrossSimON can better handle the effects of the gap between the training period and testing period. The reason behind this is that Benchmark is a forward-based prediction, which relies more on the training data (similar patterns between training data and testing data). The larger the gap, the lower the probability that the training data and testing data have similar patterns and thus the higher the probability that Benchmark has lower performance. However, CrossSimON utilizes the activity level features and network position features for making prediction (predict the counts of event types) in the target platform based on the concurrent data from the source platforms and it does not rely much on the training data. Therefore CrossSimON can better handle gap's effects on the performance of prediction and simulation.

TABLE II
Results of Simulating Reddit Events with GitHub and Twitter

Category	Measurement	Metric	CrossSimON	Benchmark
Community	comm_modularity	APE	13.677	**5.714**
	comm_num_user_actions	JSD	0.689	**0.346**
User	user_activity_dist	NRMSE	**9.779**	25.611
	user_unique_content	NRMSE	**4.271**	16.294
Content	cont_activity_ disp_geni_comm	APE	**0.980**	5.499

Note: A bold number indicates better result.

D. Discussion and Implication

Several implications are observed from the results. The gap between the training period and the testing period affects the performance of the prediction and therefore affects the performance of the simulation. The groups in different social media platforms have different activity patterns. The groups in Twitter have generally shorter lifespans compared to those in GitHub and Reddit. This will affect CrossSimON's performance in predicting bursts of activity. However, it does not affect the performance of the forward-based approach (Benchmark). This reveals the reasons that CrossSimON has relatively lower performance in predicting bursts of activity (e.g., community burstiness, community user burstiness) compared to Benchmark.

VI. Summary and Future Directions

Learning users' behavior across platforms can possibly help to detect cyber threats, but it is still a challenge due to the surge of users participating in multiple social platforms. Existing research on profile matching requires user identity information to be available, which may not be realistic. In this paper, we consider both user identities and group identities to capture the collective social behavior across platforms and develop CrossSimON to simulate online social behavior across platforms. Unlike prior works, CrossSimON considers the activity level and network position at both individual user level and group level to correlate activities across social platforms. On the one hand, CrossSimON can link user identities based on a novel probabilistic approach; on the other hand, CrossSimON presents a solution to group identity linkage across platforms. Moreover, CrossSimON is also able to simulate the information spread across multiple platforms based on the linkage of groups and users. In addition, CrossSimON leverages the advantages of prior representative approaches for linking user identities and group identities and simulates the information spread across platforms. Experimental results show that CrossSimON outperforms the Benchmark in 3 out of 5 simulation metrics. CrossSimON achieves better performance in user activity prediction. On the other hand, CrossSimON does not outperform the Benchmark in community-level metrics. The research provides new strategy for cross-platform online social network simulation, and new findings on simulating OSNs and predictive analytics for understanding online social network behavior. In the future, we will consider modeling the time lapse between different social media platforms. Also, we will consider adding more social platforms to verify the performance of CrossSimON. Finally, we will expand training data size for further improving the performance of CrossSimON.

Acknowledgment

This research was supported in part by the U.S. Defense Advanced Research Projects Agency (contract FA8650-18-C-7824) and by Fulbright Core US Scholar Fellowship (2017-19). We thank all project members, the conference editor and reviewers for their support.

References

[1] A. Smith and M. Anderson, "Social media use in 2018," https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/ [accessed in Mar. 2019].

[2] R. Wang, L. Xing, X. Wang, and S. Chen, "Unauthorized origin crossing on mobile platforms: Threats and mitigation," in *Proc. of CCS*, 2013.

[3] W. Chung, J. Liu, X. Tang, and V. Lai, "Extracting textual features of financial social media to detect cognitive hacking," in *Proc. of ISI*, 2018.

[4] F. Carmagnola and F. Cena, "User identification for cross-system personalisation," *Information Sciences*, vol. 179, no. 1-2, pp. 16–32, 2009.

[5] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migration patterns in social media," in *Proc. of AAAI*, 2011.

[6] J. Liu, H. Shen, and L. Yu, "Question quality analysis and prediction in community question answering services with coupled mutual reinforcement," *TSC*, vol. 10, no. 2, pp. 286–301, 2017.

[7] A. Narayanan and V. Shmatikov, "Myths and fallacies of "personally identifiable information"," *Communications of the ACM*, vol. 53 (6), pp. 24–26, 2010.

[8] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explorations Newsletter*, vol. 18 (2), pp. 5–17, 2017.

[9] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: A survey," *SIGKDD Explor. Newsl.*, vol. 15 (2), pp. 20–29, 2014.

[10] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *6th SNA-KDD Workshop*, 2012.

[11] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," in *VLDB*, 2014.

[12] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *IEEE 30th Symp. Security Privacy*, 2009, pp. 173–187.

[13] N. T. Ouellette, "Flowing crowds," *Science*, vol. 363, no. 6422, pp. 27–28, 2019.

[14] N. Bain and D. Bartolo, "Dynamic response and hydrodynamics of polarized crowds," *Science*, vol. 363 (6422), pp. 46–49, 2019.

[15] D. Knebel, A. Ayali, M. Guershon, and G. Ariel, "Intra- versus intergroup variance in collective behavior," *Science*, vol. 5(1):eaav0695, 2019.

[16] P. Jain, P. Kumaraguru, and A. Joshi, "@i seek 'fb.me': Identifying users across multiple online social networks," in *Proc. of WWW*, 2013.

[17] R. Roedler, D. Kergl, and G. D. Rodosek, "Content driven profile matching across online social networks," in *Proc. of ASONAM*, 2017.

[18] X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," *TKDE*, vol. 28 (2), pp. 411–424, 2016.

[19] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proc. of WWW*, Rio de Janeiro, 2013.

[20] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *ICWSM*, 2009.

[21] J. Zhang and P. S. Yu, "Multiple anonymized social networks alignment," in *Proc. of ICDM*, 2015.

[22] W. Guo, S. Wu, L. Wang, and T. Tan, "Social-relational topic model for social networks," in *Proc. of CIKM*, Melbourne, 2015.

[23] Y.-S. Cho, G. V. Steeg, E. Ferrara, and A. Galstyan, "Latent space model for multi-modal social data," in *Proc. of WWW*, Montréa, 2016.

[24] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: a full picture of relevant topics," in *Proc. of VAST*, 2014, pp. 183–192.

[25] R. K.-W. Lee, T.-A. Hoang, and E.-P. Lim, "On analyzing user topic-specific platform preferences across multiple social media sites," in *Proc. of WWW*, Perth, 2017.

[26] R. Pokharel, P. D. Haghighi, P. P. Jayaraman, and D. Georgakopoulos, "Analysing emerging topics across multiple social media platforms," in *Proc. of ACM ACSW*, Sydney, 2019.

[27] C. Taylor and D. Meldrum, "Freeway traffic data prediction using neural networks," in *Proc. of VNIS*, 1995.