NATERGM: A Model for Examining the Role of Nodal Attributes in Dynamic Social Media Networks

Shan Jiang, and Hsinchun Chen, *Fellow, IEEE*

Abstract—Social media networks are dynamic. As such, the order in which network ties develop is an important aspect of the network dynamics. This study proposes a novel dynamic network model, the Nodal Attribute-based Temporal Exponential Random Graph Model (NATERGM) for dynamic network analysis. The proposed model focuses on how the nodal attributes of a network affect the order in which the network ties develop. Temporal patterns in social media networks are modeled based on the nodal attributes of individuals and the time information of network ties. Using social media data collected from a knowledge sharing community, empirical tests were conducted to evaluate the performance of the NATERGM on identifying the temporal patterns and predicting the characteristics of the future networks. Results showed that the NATERGM demonstrated an enhanced pattern testing capability and an increased prediction accuracy of network characteristics compared to benchmark models. The proposed NATERGM model helps explain the roles of nodal attributes in the formation process of dynamic networks.

Index Terms—Social networking, graphs and networks, web mining, knowledge sharing

1 Introduction

Social media networks are emerging online networks that virtually connect individuals. These networks consist of nodes that represent individual social media users and ties that represent various relationships between the users. Examples of social media networks include online friendship networks [1], [2], following-follower networks [3], and content sharing networks [4], [5]. The relationships between the online users are often public information, which provides opportunities for using social network analysis (SNA) to better understand how and why individuals establish social connections online [6]. As a result, a growing number of studies have used SNA to examine social media networks [7], [4], [8], [5], [9].

Social media networks have two important characteristics. First, they are dynamic in nature. Network ties develop in an order, but not simultaneously. As such, relationships between individuals may change over time. Second, social media users differ in various attributes, such as gender, functional role in online communities, and reputation. As a result, social media networks are multimode networks [10], [11] and different node types exist in the network. A consequence of these two characteristics is that the seemingly same network patterns can result from different network formation processes, depending on the order in which the network ties develop. For example, Fig. 1 illustrates two processes in forming a two-star pattern. Here, we assume that the black nodes represent highly active individuals (e.g., individuals who frequently come online and leave messages) in online communities and the numbers next to network ties indicate the order in which the relationships develop. The *Pattern A* illustrates a process where highly active individuals are prioritized over others when developing relationships, while the *pattern B* illustrates the opposite tendency. If the order in which the network ties develop is ignored, we are unable to differentiate between these two patterns and understand how highly active individuals participate in the dynamic process of network formation.

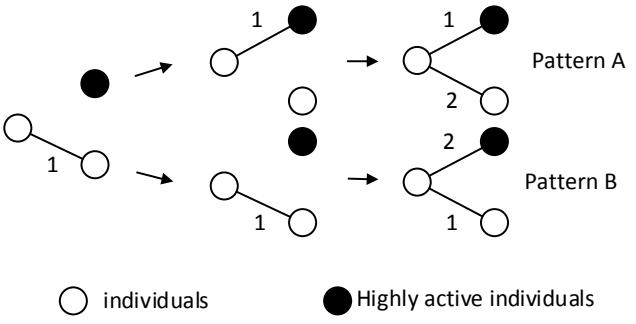


Fig. 1. Different Processes Leading to the Same Network Pattern

Differentiating between various temporal patterns is thus critical to understand the formation mechanisms of social media networks. However, current social network research usually adopts a static view of networks based on the assumption that all network ties have developed concurrently upon observation. This assumption, while contributing to simplicity and being useful for identifying static patterns of networks, leads to reduced representation of real social media networks. As a result, the ability of social network analysis to identify network patterns may be negatively affected. The problem can further re-

• *S. Jiang and H. Chen are with the Management Information Systems Department, University of Arizona, 1130 E. Helen Street, Tucson, AZ, 85721-0108. E-mail: jmly@email.arizona.edu, hchen@eller.arizona.edu.*

xxxx-xxxx/0x$xx.00 © 200x IEEE Published by the IEEE Computer Society

duce the practical value of social network analysis to understand various network phenomena in social media contexts.

In this study, we propose a novel dynamic network model, the Nodal Attribute-based Temporal Exponential Random Graph Model (NATERGM), for dynamic network analysis. NATERGM is an extension of TERGM [12] and focuses on how nodal attributes of networks affect the order in which network ties develop. The proposed model extracts nodal attributes of individuals and time information of network ties from social media networks, based on which various temporal patterns are modeled and their likelihoods of occurrence are estimated. Extending prior work [13], with empirical data we demonstrate that NATERGM provides an enhanced pattern testing capability compared to TERGM. Moreover, NATERGM is able to predict the characteristics of social media networks in future and we show that our approach outperforms TERGM-based prediction models. The major objective of this study is to provide a framework to explore, analyze, and explain the formation mechanisms of social media networks.

The remainder of the paper is organized as follows. Section 2 reviews relevant social network literature to provide the background for this research and address the need for analyzing temporal patterns in social media networks. Relevant dynamic network models are also briefly introduced and research gaps are summarized. Section 3 presents our NATERGM model. Empirical tests that demonstrate the pattern testing and network prediction performance of the model are outlined in Section 4. Results are shown and implications are discussed in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

In this section we first review recent studies examining social media networks. Then, we review emerging network models for dynamic network analysis.

2.1 Social Media Networks

Based on a theoretical conceptualization of network ties [14], four types of social media network ties have been summarized in prior research [6]. Proximity ties represent that two individuals belong to the same sub-communities (e.g., Facebook Group) or locational areas. Social relation ties represent social connections between individuals, such as virtual friendships and subscription relationships in micro-blogging sites [15], [16]. Interaction ties represent interactive behaviors between individuals, such as information exchanges via message replies [17]. Flow ties represent the movement of goods or information between network nodes, such as retweets.

Some researchers have argued that these types of ties are not necessarily decoupled, bur represent a continuum [18]. For example, proximity may further lead to social relations; interactions and flows of knowledge may occur at the same time.

Social media networks have been studied for different purposes. In general, the research objectives of these studies can be classified into three categories. The first stream of research focuses on explaining network mechanisms. This type of research aims at understanding in what conditions individuals are more likely to establish social connections online. For example, demographic homophily was found to exist in online friendship networks [19]. Students of the same gender, major, and residence area were more likely to establish social connections in Facebook friendship networks. Prior research has also found that direct reciprocity, indirect reciprocity, and preferential attachment occur very frequently in online web forums [20]. The second stream of research examines how the structure of a social media network affects the outcomes of individuals in the network. This type of research is referred to as structural capital studies [21]. For example, an examination of friendship networks in an online micro-lending platform led to discoveries that the chances of successful funding were significantly affected by the number of friendship ties and by the types of friendship [2]. Research has found that individuals in a connected network are able to predict outcomes of a given problem more accurately, compared to the cases when they are isolated [22]. Another popular research area is to partition the network into sub-graphs and detect sub-communities. These studies usually aim at identifying key groups or players in the network and understanding the characteristics of these sub-communities. For example, based on centrality and coreness measures, core groups and key members in the core group who were most active were identified in a clinical discussion forum [17]. Another study identified Twitter user clusters from following-follower networks in Twitter.com, and examined the influence of intra-group ties, inter-group ties, and intermediary ties on retweeting behaviors [3].

Previous studies focusing on community detection mainly use clustering or modularity optimization algorithms [23]. In structural capital studies, regression analysis has been frequently used to examine the relationships between network structures and individual outcomes. Dependent variables are the outcomes of network nodes, such as funding success [2] and online users' activity levels [16]. Independent variables can be various network metrics of the nodes, such as degree centrality, betweenness centrality [24], and structural holes [25]. To explain the mechanisms of network formation, network models can be used, such as the Latent Space Model [26], p1 models [27], and the Exponential Random Graph Model [28]. In social media network research, ERGM has received increased attention recently [20], [19], [29]. ERGMs are statistical models that test whether observed networks show theoretically hypothesized structural tendencies [30], [28]. These structural tendencies, or configurations, are subsets of nodes and ties in the network, reflecting certain types of network sub-structures. Examples of typical configurations can be "triangle" and "k-star" [31], [32]. In addition, nodal attributes can be incorporated in a configuration. Equation (1) specifies the expression of ERGM, where Y is a matrix of random variables representing network ties and y is its realization; η_A is a parameter corresponding to configuration A, positively related to the likelihood of configuration A to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2493549, IEEE Transactions on Knowledge and Data Engineering

occur; $g_A(y)$ is network statistics corresponding to A; κ is a normalizing constant ensuring that Pr(Y) is a probabilistic distribution.

$$\Pr(Y = y) = \left(\frac{1}{\kappa}\right) \exp\left\{\sum_A \eta_A \, g_A(y)\right\} \quad (1)$$

Given an observed network, the primary task of ERGM is to examine which configurations appeared statistically more than by chance. If a parameter η_A is estimated to be significant, it will suggest that the corresponding configuration has better chances to occur in the network, which further suggests that the corresponding effect plays an important role in the formation process of the network.

Although various analytical methods have been used to study social media networks, studies that address the dynamics of social media networks are still scarce. Only a few studies have taken into account the time information relating to when network ties are developed. For instance, Shriver et al. [16] considered the number of friendship ties at previous time points in their time series regressions. Another study analyzed the order in which retweeting links were activated in micro-blogging sites, and found that the extent to which an individual could reach other parts of the network positively affected the popularity of the content posted by that individual [33]. Overall, the dynamics of social media networks have been addressed in few prior studies. Nevertheless, dynamic network analysis is an emerging area of network research, and relevant studies have been conducted in biology, neural science, healthcare, and social science domains. We review existing dynamic network analysis approaches next.

2.2 Dynamic Network Analysis

Generally, two different approaches can be used for dynamic network analysis. Cross-sectional approaches analyze network data where time information is embedded within the network. Longitudinal approaches observe networks at multiple time points and track the evolution of networks based on comparisons [10]. Previous research has proposed various dynamic network models, including both types of approaches, for studying the dynamic process of network formation, evolution, and dissolution. We review selected dynamic network models next.

Temporal Exponential Random Graph Model (TERGM) is an extension of the ERGM for dynamic networks [34], [12], [35]. A simple TERGM model under the first-order Markov dependency can be written as:

$$\Pr(Y^t = y^t | Y^{t-1} = y^{t-1}) = \left(\frac{1}{\kappa(y^{t-1})}\right) \exp\left\{\sum_A \eta_A \, g_A(y^t, y^{t-1})\right\} \quad (2)$$

Note that the major difference between (1) and (2) is the specification of network statistics for each temporal pattern A, which is now determined by network realizations in multiple observational time points (observed at t and t-1 in this case). Given multiple observations, TERGM can be used to test whether a certain temporal pattern is more likely to occur than by chance. For example, as illustrated in Fig. 2, three different temporal patterns can be derived from a transitivity pattern, depending on the order in which the three ties develop. Compared to the conventional ERGM where only a tendency for transitivity can be tested, TERGM differentiates between three different dynamic patterns of network ties formation which all finally lead to the same transitivity structure in (a). TERGM can further test the likelihood of each temporal pattern to occur.

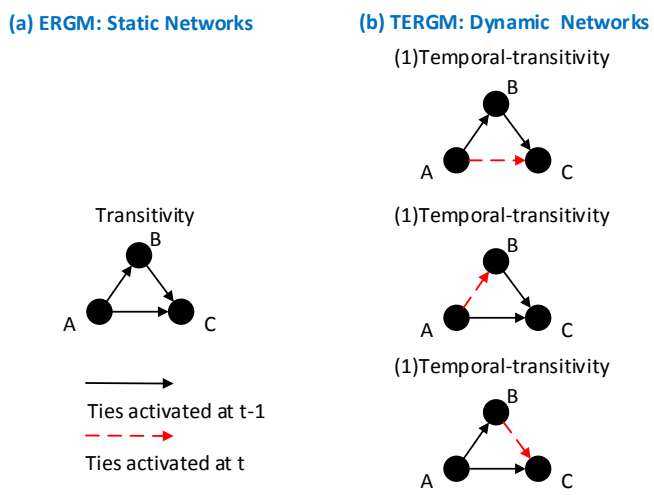


Fig. 2. Three Different Temporal Patterns Derived From Transitivity

In addition to the transitivity in this example, TERGM can also include network configurations of many other types such as temporal stability and temporal reciprocity [12], [36]. TERGM can also be applied to cross-sectional data if time duration information for network ties is provided. However, none of the TERGM research has considered how nodal attributes can affect the order in which network ties develop.

Separable Temporal Exponential Random Graph Model (STERGM) separates TERGM into a formation model and a dissolution model, thereby modeling not only the temporal patterns of network formation, but also the temporal patterns of network dissolution [37], [38], [36]. STERGM addresses the concern that some existing network ties might disappear over time, such as a broken friendship, for example. STERGM identifies new connections and dissolved ties by comparing networks at multiple time points. A variant of STERGM for cross-sectional data is also proposed for the case when longitudinal data is unavailable [38].

Hidden Temporal Exponential Random Graph Model (HTERGM) is a model that combines TERGM with hidden Markov models [34]. It assumes that (1) network structure at time t, Y_t, is dependent on the structure of the network in the previous time point Y_{t-1}, and (2) nodal attributes of the network, x_t, are dependent on the network structure Y_t. It further assumes that only nodal attributes are observable, while network structures are hidden states. The major aim of HTERGM is to estimate the transition probabilities $P(Y_t | Y_{t-1})$ and emission matrices $\Lambda = P(x_t | Y_t)$ so that hidden network structures can be inferred given time series of nodal attributes $x_1, x_2, \ldots, x_t$. However, HTERGM does not explain how nodal attributes affect the formation process

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2493549, IEEE Transactions on Knowledge and Data Engineering

of networks.

Temporally Randomized Reference Models (TRRM) investigates the dynamic characteristics of networks by comparing observed networks with an ensemble of temporally randomized networks [39], [40], [41]. Temporal randomization generates new networks by rewiring ties in the original networks or changing time information associated with the ties. Typical randomization methods include randomized edges, randomly permutated times, random times, edge randomization, and time reversal [40]. Fig. 3 shows examples of randomized edges and randomly permutated times. By comparing original networks with temporally randomized networks, key dynamic characteristics of original networks can be understood. For example, Holme [39] compared e-mail networks with their temporally randomized samples and found that in general the average time it took to pass information between network nodes is longer in the original email networks.

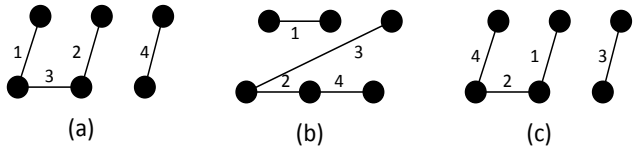


Fig. 3. Network Temporal Randomization with (a) an original network with numbers indicating the order of tie activation; (b) a randomized network by iteratively rewiring network ties among four selected nodes; and (c) another randomized network by permuting the time associated with ties.

Latent space models [26] assume that each node in a network is associated with a latent position in a low dimensional space. The probabilities of tie occurrences are determined by the distances between nodes in the latent space. The latent space model estimates the parameters associated with latent positions based on the observed networks. The estimated model can be used to visualize a spatial representation of network relationships [26], [42]. Dynamic Latent Space Model (DLSM) is an extension of the latent space model and allows the latent positions to change over time [43], [44].

2.3 Research Gaps

Based on the prior literature, several research gaps can be identified. First, social media networks are dynamic in nature. However, little research has explained the mechanisms of network formation with a dynamic perspective. Dynamic network analysis has been frequently used to detect communities from networks [10], [11], but not to explain the mechanisms of network formation. Most network mechanisms studies focused on identifying static network patterns, but did not explain how these patterns developed dynamically. Second, emerging network research has given rise to various approaches for examining temporal networks and has suggested that the order of network ties is an important aspect of network dynamics [12], [40], [33]. Recent TERGM models examine different dynamic patterns of network tie formation in dyadic and triadic relationships when all the nodes are considered to be of the same type. STERGM additionally examines the order in which network ties dissolve. However, none of the existing models explain even more complex patterns created by the interactions of network tie order and nodal attributes. We need a model to carefully examine such interactions in order to understand how nodal attributes affect the order in which network ties develop. In addition, network prediction has been an under-studied research area [45]. Although prior research has helped identify dynamic network patterns, little has been done to predict future networks based on the identified patterns.

3 Nodal Attribute-based Temporal Exponential Random Graph Model

The proposed NATERGM focuses on how nodal attributes of networks affect the order in which network ties develop. Because the order of network ties needs to be tracked accurately, NATERGM examines cross-sectional network data with time information for network ties. Figure 4 presents the framework of NATERGM. The major components include network extraction, temporal pattern analysis, and network prediction. In the network extraction step, social connections are identified between individuals in social media, along with the timestamps of these relationships and nodal attributes of the individuals. Temporal patterns of the networks are modeled, and the likelihood of each pattern is estimated in the temporal pattern analysis step. Based on the estimated model, new networks are simulated and compared to the original network to evaluate how effectively the model can predict future networks. We explain each component in the following subsections.

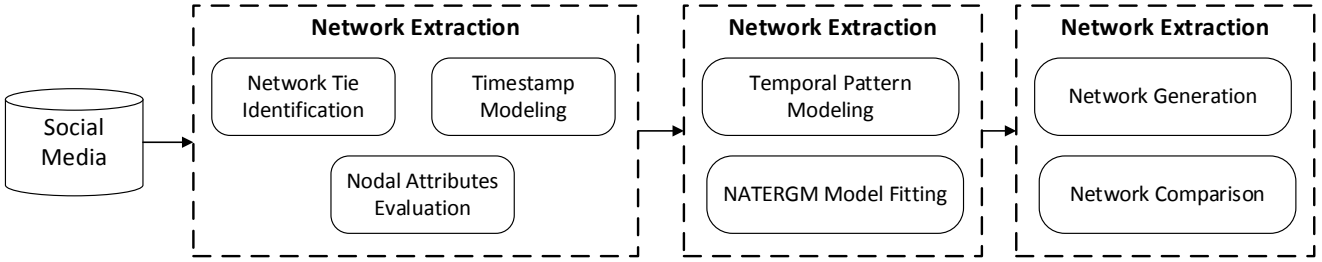


Fig. 4. NATERGM Framework

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2493549, IEEE Transactions on Knowledge and Data Engineering

3.1 Network Extraction

First, network ties are extracted from social media based on relationships between online users. Among the various types of social media network ties summarized by Kane et al. [6], the interaction/flow and social relation ties are the ones that are the most dynamically established (i.e., these ties are often associated with timestamps). Different types of network ties can be identified depending on specific social media contexts. For example, directed interaction/flow ties can be established if an individual sends greetings to another individual; undirected social relation ties can be established if two individuals become friends by using friending functions provided in social media platforms. After identifying network ties between all possible pairs of individuals, a network with N nodes is represented by a matrix $Y=[Y_{ij}]$, (i, j =1, 2,…N). For undirected networks, $Y_{ij}=1$ if a tie exists between nodes (i.e., individuals) i and j, and $Y_{ij}=0$ otherwise. For directed networks, $Y_{ij}=1$ if a tie starts from i and ends at j, and $Y_{ij}=0$ otherwise.

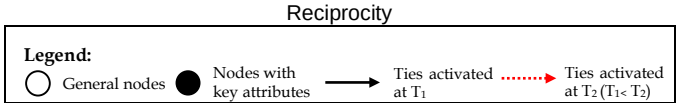
For timestamp modeling, we use T_{ij} to represent the time when each network tie (i, j) is established. A matrix $\boldsymbol{T}=[T_{ij}]$, (i, j =1,2,…N) records the timestamps for all network ties and can be used to model the order of network ties. For example, if $T_{12}<T_{21}$, it would represent a process where node 1 sent out a tie to node 2 first, and then received a tie from the node 2 in return.

Nodal attributes of individuals can be evaluated using different approaches. Prior studies have characterized individual social media users based on three types of features. Platform-based features refer to individual attributes that are directly provided by social media platforms. For example, registered users are often associated with usernames while an unregistered user is represented by a "visitor" tag or an IP address in the name space. Some social media platforms also assign functional roles to users such as members or administrators. This type of information can be directly used as nodal attributes of individuals. Textual features refer to attributes that are inferred by texts posted by the individuals. Social media users typically leave many textual traces, such as private messages and message postings. Various characteristics of social media users can be evaluated based on these texts, such as general opinions, writing proficiency, and topics of interests. Social network features refer to individual attributes that are inferred by their connections or positions in the network. Social relations between individuals in part reflect their personality, status, and roles. For example, an individual who is linked with many others is expected to have a high level of popularity compared to others who have fewer connections. Such information can thus be used as nodal attributes of individuals. After evaluating the nodal attributes of individuals, they are represented by a vector $\boldsymbol{X}=(x_1, x_2, ..., x_N)$.

3.2 Temporal Pattern Analysis

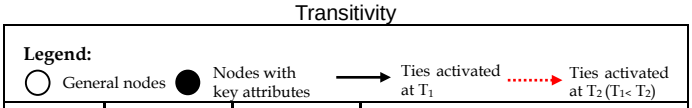
To model temporal patterns, the nodal attributes and timestamps of network ties are used to represent various temporal patterns regarding the dynamics of network formation. By taking into account the order in which network ties develop, common static network patterns such as reciprocity, k-star, transitivity, and cyclicity can have different temporal variations. Tables 1 to 5 list examples of temporal patterns for directed networks. White nodes represent individuals in general and black nodes represent individuals with key nodal attributes (e.g., highly active individuals). Dashed arrows represent network ties that developed after solid ones.

TABLE 1.
NATERGM Temporal Patterns for Directed Networks: Reciprocity



Static Pattern	Temporal Pattern	Illustration	Hypothesis
reciprocity	feedback		Nodes with some attribute have a high tendency to receive feedback.
	response		Nodes with some attribute have a high tendency to respond to incoming ties.

TABLE 2.
NATERGM Temporal Patterns for Directed Networks: Transitivity



Static Pattern	Temporal Pattern	Illustration	Hypothesis
transitivity	bridge		Nodes with some attribute have a high tendency to bridge new relationships between others.
	co-supporting		If two nodes are supporting a common node with some attribute, they have a high tendency to build a new relationship.
	co-supported		If two nodes are supported by a common node with some attribute, they have a high tendency to build a new relationship.
	remarked-supporter		A node with some attribute has a high tendency to receive attention from another node, if both co-support a common node.
	remarked-supported		A node with some attribute has a high tendency to receive attention from another node, if they are supported by a common node.
	remarking-supporter		A node with some attribute has a high tendency to pay attention to another node, if they co-support a common node.
	remarking-supported		A node with some attribute has a high tendency to pay attention to another node, if they are supported by a common node.
	follow-up		A node with some attribute has a high tendency to pay attention to another node, if a third node bridges their relationship.
	reference		A node with some attribute has a high tendency to receive attention from another node, if a third node bridges their relationship.

As can be seen from the table, the temporal patterns modeled by NATERGM provide an extended hypotheses testing capability about network formation compared to static patterns. In particular, these temporal patterns can be used to examine the roles of nodal attributes in determining the order of network ties. For example, assuming

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2493549, IEEE Transactions on Knowledge and Data Engineering

TABLE 3.
NATERGM Temporal Patterns for Directed Networks: K-out-star

Legend: General nodes; Nodes with key attributes; Ties activated at T_1; Ties activated at T_2 ($T_1 < T_2$)

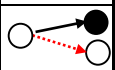
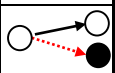
Static Pat-	Temporal Pattern	Illustration	Hypothesis
k-out-star	prioritization		Nodes with some attribute have a high tendency to be prioritized when forming relationships.
	de-prioritization		Nodes with some attribute have a high tendency to be de-prioritized when forming relationships.

TABLE 4.
NATERGM Temporal Patterns for Directed Networks: K-in-star

Legend: General nodes; Nodes with key attributes; Ties activated at T_1; Ties activated at T_2 ($T_1 < T_2$)

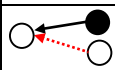
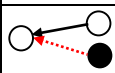
Static Pat-	Temporal Pattern	Illustration	Hypothesis
k-in-star	initiative		Nodes with some attribute have a high tendency to take the initiative in multi-actor relationships.
	laziness		Nodes with some attribute have a high tendency to hold off in multi-actor relationships.

TABLE 5.
NATERGM Temporal Patterns for Directed Networks: Cyclicity

Legend: General nodes; Nodes with key attributes; Ties activated at T_1; Ties activated at T_2 ($T_1 < T_2$)

Static Pat-	Temporal Pattern	Illustration	Hypothesis
cyclicity	reversed-refenrence		A node with some attribute has a high tendency to receive attention from another node. if a third node
	reversed-follow-up		A node with some attribute has a high tendency to pay attention to another node, if a third node bridges their relationship reversely.
	reversed-bridge		Nodes with some attribute have a high tendency to reversely bridge new relationships between others.

that we are interested in the role of highly active individuals in developing message flows in social media, the static reciprocity pattern would only model a tendency for two individuals (at least one of them being highly active) to exchange messages. In comparison, if we observed many "feedback" patterns in the network, it would suggest a tendency for highly active individuals to receive returning messages after they sent out messages first; if we observed many "response" patterns, it would suggest a tendency for highly active individuals to respond to others' incoming messages. Although both "feedback" and "response" patterns finally lead to the same "reciprocity" pattern, they model two distinct dynamic processes. In a similar way, NATERGM extends other static patterns (i.e., k-star, transitivity, and cyclicity) to their temporal variations by considering the possible order of network ties, which provides richer insight about the dynamic process of network formation.

Given the list of temporal patterns in Tables 1 to 5, the major objective of NATERGM is to test which of these temporal patterns are more likely to be observed than to occur by chance in a network. The NATERGM model can be written as:

$$\Pr(Y = y|\boldsymbol{\eta}) = \left(\frac{1}{\kappa}\right) \exp\left\{\sum_{a \in A} \eta_a \, g_a(y, \boldsymbol{T}, \boldsymbol{X})\right\} \quad (3)$$

In (3), A is a set of temporal patterns to be tested, $\boldsymbol{\eta} = [\eta_a]$ is a vector of parameters representing the strength of each temporal pattern's effect in network formation, and κ is a scaling parameter to ensure (3) is a probability distribution. $g_a(\bullet)$ is the network statistic of temporal pattern a, evaluated with network y, timestamp matrix T, and vector of nodal attributes X. Table 6 provides definition of $g_a(\bullet)$ for each temporal pattern listed in Tables 2 to 5, with the assumption that nodal attributes are binary or categorical. $I()$ is an indication function that takes the value 1 if and only if the expression inside results in TRUE values. For categorical attributes, $I(X_j)$ takes the value 1 if node i belongs to the desired category in X. For cases when nodal attributes are continuous variables, $I(X_i)$ is replaced by the value of X_i.

TABLE 6.
Specification of NATERGM Terms
(Directed Network, Binary or Categorical Attributes)

NATERGM Term	Network Statistic
reciprocity	
feedback	$g_F(y,T,X) = \sum_{i \neq j} y_{ij} \cdot y_{ji} \cdot I(X_i) \cdot I(T_{ij} < T_{ji})$
response	$g_R(y,T,X) = \sum_{i \neq j} y_{ij} \cdot y_{ji} \cdot I(X_i) \cdot I(T_{ij} > T_{ji})$
2-out-star	
prioritization	$g_P(y,T,X) = \sum_{i \neq j \neq k} y_{ki} \cdot y_{kj} \cdot I(X_i) \cdot I(T_{ki} < T_{kj})$
deprioritization	$g_D(y,T,X) = \sum_{i \neq j \neq k} y_{ki} \cdot y_{kj} \cdot I(X_i) \cdot I(T_{ki} > T_{kj})$
2-in-star	
initiative	$g_I(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ik} < T_{jk})$
laziness	$g_L(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ik} > T_{jk})$
transitivity	
bridge	$g_B(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{jk} > T_{ji}, T_{ik})$
cosupporting	$g_{CS}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{jk} > T_{ji}, T_{ki})$
cosupported	$g_{CSD}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{jk} > T_{ij}, T_{ik})$
remarked-supporter	$g_{RS}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ji} > T_{ik}, T_{jk})$
remarked-supported	$g_{RSD}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ji} > T_{ki}, T_{kj})$
remarking-supporter	$g_{RMSR}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ij} > T_{ik}, T_{jk})$
remarking-supported	$g_{RMSD}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ij} > T_{ki}, T_{kj})$
follow-up	$g_{FU}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ij} > T_{ik}, T_{kj})$
reference	$g_{REF}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ji} > T_{jk}, T_{ki})$
cyclicity	
reversed_reference	$g_{RREF}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ki} > T_{ij}, T_{jk})$
reversed_followup	$g_{RFU}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{ij} > T_{ki}, T_{jk})$
reversed_bridge	$g_{RB}(y,T,X) = \sum_{i \neq j \neq k} y_{ik} \cdot y_{ij} \cdot y_{jk} \cdot I(X_i) \cdot I(T_{jk} > T_{ij}, T_{ki})$

The likelihood of occurrence for each temporal pattern can be assessed by estimating the parameters $\boldsymbol{\eta}$. If a parameter is positive and significant, it indicates that the corresponding temporal pattern appears more frequently than by chance in the network. For parameter estimation, the

Markov Chain Monte Carlo (MCMC) method is used, following prior ERGM literature [46]. The procedure is modified to adapt to temporal settings. The overall procedure is illustrated in Figure 5.

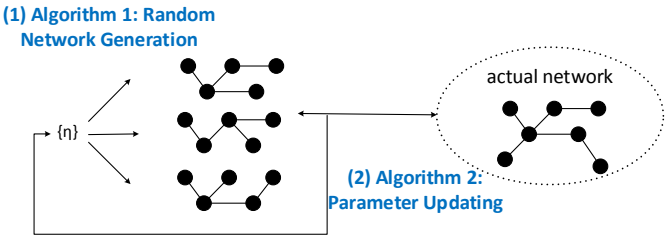


Fig. 5. NATERGM Parameter Estimation Procedure

In general, the model fitting procedure iteratively generates random networks based on the given set of parameters and updates the parameters based on the difference between the generated networks and the observed network. For a given set of parameters $\boldsymbol{\eta} = [\eta_a]$, **Algorithm 1** is used to generate random networks on a given set of nodes.

Algorithm 1. NATERGM Random Network Generation
Initialize network as $Y = Y^{(t=0)}$
repeat until maximum rounds of iterations are made
for each element Y_{ij} in $Y^{(t)}$:
change the value of Y_{ij} based on the conditional distribution defined by

$$logit\{\Pr(Y_{ij} = 1 | Y_{kl} = y_{kl}\ for\ all\ (k,l) \neq (i,j))\} = \boldsymbol{\eta}^T \left(\boldsymbol{g}\left(y^{(ij1)}, T_{Gibbs}, X\right) - \boldsymbol{g}\left(y^{(ij0)}, T_{Gibbs}, X\right)\right)$$

end for
t←t+1
return Y(t)

$y^{(ij1)}$ and $y^{(ij0)}$ are matrices that only differ in element Y_{ij}, taking 1 or 0 respectively, $\boldsymbol{g} = \{g_{a1}, g_{a1}, g_{a1}, \dots g_{aM},\}$ and definitions for $\boldsymbol{g}(\bullet)$ can be found in Table 6. T_{Gibbs} is a timestamp matrix where $T_{ij} > T_{kl}$ for all $(k,l) \neq (i,j)$. It assumes that the stochastic process Y(t) develops over time. Based on the logit value calculated (say, q), the probability that a network tie changes its value to 1 can be calculated as p=exp(q)/(1+exp(q)), based on which we use Monte Carlo method to draw a new value for Y_{ij}.

Given the random network generation procedure, **Algorithm 2** is used to estimate parameter values. It calculates the differences for a set of network statistics between generated networks and the actual network, and use the differences to adjust the parameters used to generate the networks.

Algorithm 2. NATERGM Parameter Updating
initialize $\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)}$
repeat from n=0:
generate K networks (y_1, y_2 …y_K) independently based on $\boldsymbol{\eta}^{(n)}$ and **Algorithm 1**
define

$$\bar{\boldsymbol{g}} = \left(\frac{1}{K}\right)\sum_{k=1}^{K}[p_k^{(n)}\boldsymbol{g}(1 - y_k^{(n)}) + (1 - p_k^{(n)})\boldsymbol{g}(y_k^{(n)}) - \boldsymbol{g_0}]$$

and

$$\boldsymbol{D_0} = diag\{\left(\frac{1}{K}\right)\sum_{k=1}^{K}[p_k^{(n)}\boldsymbol{g}^T(1 - y_k^{(n)})\boldsymbol{g}(1 - y_k^{(n)}) + (1 - p_k^{(n)})\boldsymbol{g}^T(y_k^{(n)})\boldsymbol{g}(y_k^{(n)}) - \bar{\boldsymbol{g}}^T\bar{\boldsymbol{g}}]\}$$

calculate

$$\boldsymbol{Z}^{(n)} = (Z_1^{(n)}, Z_2^{(n)}, \dots, Z_K^{(n)}),$$
$$Z_K^{(n)} = p_k^{(n)}(y)(1 - y_k^{(n)}) + (1 - p_k^{(n)}(y))\boldsymbol{g}(y_k^{(n)}) - \boldsymbol{g_0}]$$

where

$$p_k^{(n)}(y) = \frac{\exp(\boldsymbol{\eta}^{T(n)}\boldsymbol{g}(1 - y_k))}{\exp(\boldsymbol{\eta}^{T(n)}\boldsymbol{g}(1 - y_k) + \exp(\boldsymbol{\eta}^{T(n)}\boldsymbol{g}(y_k)}$$

update $\boldsymbol{\eta}^{(n+1)}$ using Robins-Monro Algorithm $\boldsymbol{\eta}^{(n+1)} = \boldsymbol{\eta}^{(n+1)} - s_n\boldsymbol{D_0^{-1}}\boldsymbol{Z}^{(n)}$, n←n+1
until convergence criterion is met

s_n is a sequence of positive numbers converging to 0. In this study we used $s_n = 2\exp(n)/10$, as suggested in prior research [46]. For convergence criterion, we also used the t-ratio methods in [46].

3.3 Network Prediction

After estimating the parameters in NATERGM, the fitted model can be used to predict the characteristics of future networks with the following procedures.

Based on the actual network observed at time point t-1, NATERGM parameters $\boldsymbol{\eta_{t-1}}$ are estimated. A number (=K) of networks at time point t are then simulated based on the parameters $\boldsymbol{\eta_{t-1}}$ using **Algorithm 1**. However, network at the time point t-1 is used as the initial network, instead of a randomly initialized network.

Each generated network at time point t does not necessarily look exactly like the actual network at time point t. However, global network statistics averaged over K generated networks should resemble those of the actual network. An assumption made here is that global network property does not change dramatically in a short term [55], and thus a network model estimated at time t-1 should be able to generate networks that are also similar to networks in time t in terms of global network statistics. Moreover, the parameters $\boldsymbol{\eta^{t-1}}$ used for network generation in the proposed model are related to the tendency of corresponding temporal patterns, which should be reflected gradually over time in networks. Therefore, we use the similarity between generated networks with the actual network in the next time period to evaluate the prediction performance.

In order to evaluate how close the generated networks are to the actual network in the next period, we calculate the absolute difference (AD) for each network statistic a′ A′ at prediction period t:

$$AD_{a'}^t = |g_{a'}(y_0^t) - \left(\frac{1}{K}\right)\sum_{k=1}^{K} g_{a'}(y_k^t)| \quad (4)$$

where y_0^t is the observed network at t, and y_k^t (k=1,2,…, K) is the k-th generated network based on the fitted model at t. Small difference would indicate that the estimated model predicts the network well.

4 Research Design

In order to evaluate the performance of NATERGM, we conducted two empirical tests. The first test focused on the pattern testing capability of NATERGM. The second test

focused on how accurately our model can predict the characteristics of future networks. This section describes the research test-bed used for empirical study and outlines the two experiments.

4.1 Research Test-bed

Social media data were collected from WikiAnswer.com, which is a large online knowledge sharing community. Community members can ask questions about any topics, and answer others' questions as well. Open questions go through the hands of many contributors, some of whom directly provide answers, while others edit the posted answers in terms of content, language, or format. Finally, the questions and answers are organized into Q&A entries that can be accessed by all community members including the questioners. It is a good test-bed for testing NATERGM because its wiki-based "answer history" system allows us to see how members in this community develop social connections when seeking help and answering questions. We established a directed tie from member A to member B if A answered a question from B. Therefore, network ties represent knowledge flows in this community and as a result, knowledge diffusion networks were extracted from social media. The test-bed also allows for identifying timestamps associated with these network ties.

Since many questions require contributors to have relatively deep knowledge in a field to answer, members in WikiAnswer.com form specialized sub-communities to handle questions that belong to similar topics. In this study, we focused on three sub-communities: diabetes, online shopping, and real estate, based on the popularity of these topics and the number of relevant Q&A entries in the community. Furthermore, we observed that a great number of members were inactive and only participated in knowledge sharing activities very limitedly (e.g., only made a change to capitalization once). Since we were interested in the most representative members in the community, we restricted analysis within members who asked questions, provided the first answers, or made significant content change (this type of contribution is separately classified in WikiAnswer.com) to answers more than once. The data collection statistics are shown in Table 7. In the resulting networks from WikiAnswer.com, each user made connections with others (by providing or receiving answers) 1.8 times on average. The network density ranged from 0.06% to 0.2% for the three sub-community networks, suggesting that these networks are quite sparse. Overall, the degree (number of connections) of individual users followed the power law distribution, with the highest degree as 14.

TABLE 7.
Data Collection from WikiAnswer.com

Community	# of Q & A Entries	Total # of Members	# of Members for Analysis	Time Span
Diabetes	27,349	70,164	2,499	2008-2013
Online Shopping	14,111	39,631	1,058	2008-2013
Real Estate	10,725	36,517	963	2008-2013

4.2 Pattern Testing

In order to show the enhanced pattern testing capability of NATERGM, we compared it with TERGM and used both models to explain network formation. TERGM was chosen as the baseline model because it is also capable of modeling the order in which network ties develop. However, it does not explain what roles nodal attributes play in determining the order of network ties.

NATERGM and TERGM used for pattern testing included different sets of model terms. Baseline model terms included arc, reciprocity, 2-out-star, 2-in-star, transitivity, and cyclicity. For all the significant terms in the baseline model, they were further extended into corresponding temporal patterns in Tables 1 to 5 and tested using NATERGM. For terms that were not significant in the baseline model, their extensions need not be tested because the extended temporal patterns are subset of their corresponding root patterns by additionally considering nodal attributes. Therefore, if a root pattern does not frequently appear in the network, its extension would not become significant either. We evaluated three types of nodal attributes for pattern testing with NATERGM.

For platform-based features, we identified registered members and unregistered visitors (attr=reg). Although WikiAnswer.com does not mandate a new user to register for posting or answering questions, registered members can accumulate "trust points" and obtain honor badges based on their contributions over time. We expected that registered members might have more commitment to the community and more willingness to contribute than unregistered anonymous visitors, thereby showing different behaviors from unregistered visitors when developing network ties. The attribute was measured as a binary variable, where reg = 1 if the member was registered, and reg=0 otherwise.

For textual features, we evaluated writing proficiencies (attr=pro) for community members. Writing proficiency reflects an individual's level of literacy, expertise, and educational background [47, 48]. We expected that members with high levels of writing proficiency would contribute significantly to the online knowledge sharing communities. Hence, we were interested in understanding what roles these members would play in the formation process of the dynamic knowledge diffusion networks. Prior linguistic studies have suggested that writing proficiency can be assessed based on various factors [49], [50]. We employed text mining techniques to evaluate the following five metrics of each member in WikiAnswer.com. First, the average length of a member's answers was evaluated because it reflects the depth of the member's knowledge about the problem [51]. Second, T-unit (a single main clause + other subordinate clauses) is an index of syntactic complexity [52] and reflects the member's effectiveness in organizing words in sentences. We used the number of words per T-unit to evaluate this metric. Third, lexical richness measures the extent to which different words are used in texts. We used Hapax Legomena and Hapax Dislegomenna to evaluate lexical richness, following prior studies [48], [47]. Fourth, objectivity reflects whether answers

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2493549, IEEE Transactions on Knowledge and Data Engineering

are provided without biases. Objectivity of a member was evaluated based on the percentage of sentences that are classified as "objective" by the Opinion Finder (OF) System [53], [54] of all sentences posted by that member. Fifth, we used one minus percentage of misspellings (i.e., percentage of correct spellings) to measure the readability of texts for each member. Finally, these metrics were normalized to [0, 1] and averaged to represent the member's writing proficiency.

For social network features, we evaluated out-degree centrality (attr=odc) of each member. The out-degree centrality of a member in this knowledge sharing community reflects how actively the member helps answering others' questions. Individuals with high out-degree centrality should play a critical role in the community, and thus we were interested in understanding how they would impact network formation dynamically. The out-degree centrality of a member was measured based on the number of outgoing ties associated with him or her, and was also normalized to [0, 1], so that the range of attribute values are consistent across all three types of nodal attributes.

4.3 Network Prediction

To evaluate the performance on network prediction, we compared the prediction results of NATERGM and TERGM. For NATERGM, the prediction procedure followed the procedure described in Section 3.3. For TERGM, the model was trained using the previous two time points t-2 and t-1 in order to predict network at time point t. As for network statistics to be compared, we focused on comparing the degree distribution of simulated networks and the actual network following previous studies [13, 29], and the set of network statistics A' included standard deviation and skewness of in-degree and out-degree distributions of nodes. The major reason why we selected them for comparison is that we believe in social media networks, the degree distribution best reflects how people are connected and participate in social interactions. Especially in WikiAnswer.com, the existence and distribution of highly active users and less active users is a key characteristics in the network. Prediction was conducted for each month during 2008-2013. Absolute difference vectors (AD_a^1, AD_a^2 ,…, AD_a^{72}) were obtained for both the baseline model and NATERGM.

The Wilcoxon signed rank test was used to validate that the prediction errors of the NATERGM were statistically lower than that of the baseline model. The procedure first calculated the differences of absolute prediction errors between the two models ("TERGM" minus "NATERGM"), and ranked the pairs according to the absolute values of the differences (from lower to higher). It then calculated the sum of the ranks where the differences were positive, as shown below.

$$W^+ = \sum_{t=1}^{T} I(AD^t{}_{baseline} - AD^t{}_{NATERGM} > 0) \cdot R_t \qquad (5)$$

T is the total number of periods for comparison, $I()$ is an indication function that the prediction error of baseline is greater than that of the NATERGM, and R_i is the rank of the pair.

The Z statistic was calculated as $Z = (|W^+ - \mu_W| - 0.5)/\sigma_W$, where $\mu_W = n(n+1)/4$, and $\sigma_W = \sqrt{n(n+1)(2n+1)/24}$. High Z values would indicate that the differences of prediction errors between the two models were statistically significant.

5 Results and Discussions

5.1 Pattern Testing Results

For each of the selected nodal attributes including registration status (reg), writing proficiency (pro), and out-degree-centrality (odc), we tested how the attribute affected the dynamic process of network formation by fitting the NATERGM model to networks extracted from each of the three sub-communities in our test-bed. Table 8 shows the estimated parameters for the baseline model and NATERGM.

TABLE 8. Parameter Estimates

Model	Model Terms	Estimates (S.E.)					
		Diabetes		Online Shopping		Real Estates	
TERGM	**arc**	**-5.11*** **	**(0.72)**	**-4.85*** **	**(0.78)**	**-5.07*** **	**(0.63)**
	reciprocity	0.59	(1.01)	0.41	(1.23)	0.59	(0.97)
	2-out-star	**2.51*** **	**(0.40)**	**2.34*** **	**(0.44)**	**2.74*** **	**(0.38)**
	2-in-star	**2.56***	**(0.85)**	**2.45***	**(0.79)**	**2.63** **	**(0.73)**
	transitivity	1.31	(1.05)	1.16	(0.72)	1.28	(1.10)
	cyclicity	-0.09	(0.72)	0.41	(0.46)	-0.17	(0.24)
NATERGM	**arc**	**-5.26*** **	**(0.81)**	**-5.16*** **	**(0.62)**	**-5.33*** **	**(0.59)**
(NA=reg)	2-out-star extended						
	[reg]-prioritization	**2.66*** **	**(0.37)**	**2.35*** **	**(0.45)**	**2.78** **	**(0.63)**
	[reg]-deprioritization	**2.23***	**(1.01)**	**2.39***	**(0.89)**	**2.26***	**(0.94)**
	2-in-star extended						
	[reg]-initiative	**3.43** **	**(0.97)**	**3.70** **	**(1.01)**	**2.83***	**(0.95)**
	[reg]-laziness	1.51	(1.01)	1.23	(1.12)	1.58	(1.57)
NATERGM	**arc**	**-4.72*** **	**(0.73)**	**-5.05*** **	**(0.67)**	**-5.27*** **	**(0.81)**
(NA=pro)	2-in-star extended						
	[pro]-laziness	**2.68*** **	**(0.58)**	**2.99** **	**(0.70)**	**3.22** **	**(0.73)**
NATERGM	**arc**	**-5.94***(1.01)**		**-5.41*** **	**(0.79)**	**-5.72*** **	**(0.93)**
(NA=odc)	2-in-star extended						
	[odc]-initiative	**3.87*** **	**(0.61)**	**3.46*** **	**(0.53)**	**3.81*** **	**(0.73)**

NOTE: *: p<0.05, **: p<0.01, ***: p<0.001. Significant variables are marked in boldface.
NA=Nodal Attribute.

The estimated parameter values are log-odds of the ties of corresponding patterns forming in the network. These values are positively related to the probabilities that corresponding patterns would occur. For example, the value of estimated parameter for the configuration "2-out-star" in Table 5 is positive (2.51). Then the conditional log-odds of a directed tie adding a "2-out-star" pattern is -5.11 + 2.51 = -2.60 (because the tie automatically adds an "arc" as well). The probability that such a tie would develop in the network is exp(-2.60)/(1 + exp(- 2.60)) = 0.065, which is greater than the probability that a directed tie would develop between any pair of nodes, which is exp(-5.11)/(1 + exp(-5.11)) = 0.006.

For the nodal attribute of registration status, similar results were observed in all three sub-communities. In the baseline model, we observed that "2-out-star"and "2-in-star" terms were significant. It indicates that members in the community were likely to help more than one peer over time by providing answers to their questions, and questioners were also likely to receive answers from multiple contributors over time. When these patterns were further extended to their corresponding temporal terms in NATERGM, additional insights could be obtained. For example, when temporal 2-out-star patterns were considered, we found that both "[reg]-prioritization" and "[reg]-deprioritization" terms were positive and significant. This suggests that when a member helped others, he or she did not prioritize registered members over unregistered visitors in terms of time. In other words, everyone in the community was treated equally in terms of receiving answers. However, for temporal 2-in-star patterns, only the "[reg]-initiative" term was significant. It suggests that when a member receives answers from multiple contributors, the answers were likely to come from registered members first. An explanation is that registered members may have more commitment and want to establish a good image in the community. As a result, registered members may try to help others as soon as possible. Using NATERGM terms, we saw that the proposed model ws able to reveal how registered members affect the dynamic process of knowledge diffusion.

When writing proficiency was considered, none of the extended 2-in-star terms were significant. However, when extending the 2-out-star terms, we found that only the term "[pro]-laziness" was positive and significant. This suggests that when a member receives answers from multiple contributors, answers from members with good writing proficiency tended to arrive later. A possible reason is that members with good writing proficiency try to provide well-constructed, unbiased, and helpful answers to questioners. They may take some time to do research before providing answers, which delays the time when they delivered answers to questioners. Similar results were observed in all-three sub-communities. Using NATERGM terms, we saw that the proposed model was able to reveal how the writing proficiency of members affect the dynamic process of knowledge diffusion.

As for out-degree centrality, none of the extended 2-out-star terms were significant. However, when temporal 2-in-star patterns were considered, we found that only the "[odc]-initiative" term was positive and significant. It suggests that when a member receives answers from multiple contributors, answers from members with high out-degree-centrality were likely to arrive early. High out-degree centrality of a member indicates that he or she frequently helps others, reflecting the member's high level of activity and willingness to help. Consequently, members with high out-degree centrality tried to help others as early as possible. Again, using NATERGM terms, we saw that the proposed model was able to reveal how the out-degree centralities of members affect the dynamic process of knowledge diffusion.

In sum, by comparing the pattern testing results of TERGM and NATERGM, we could always obtain additional insights about how nodal attributes of social media users affect the dynamic process of network formation. Therefore, NATERGM has an enhanced pattern testing capability compared to the benchmark network model.

5.2 Network Prediction Results

Table 9 shows the results of Wilcoxon signed rank tests for the prediction errors of NATERGM and TERGM. The numbers are Z-statistics and the asterisks indicate the level of significance. Results show that the prediction errors obtained by NATERGM were statistically lower than that of TERGM for all of the four selected network statistics. Therefore, NATERGM was able to predict the characteristics of networks more accurately than TERGM in terms of the networks' degree distribution in our setting. This provides a positive case that NATERGM has the potential to make better prediction than TERGM. However, since other factors such as the length of prediction periods were not controlled for both models, additional test cases are needed to make this statement more convincing.

TABLE 9.
Wilcoxon Signed Rank Tests for Prediction Errors

Community	Test statistics			
	std. in-degree	std. out-degree	skewness in-degree	skewness out-degree
Diabetes	6.26***	5.73***	4.44**	4.92***
Online Shopping	5.12***	4.86 ***	4.49**	4.77**
Real Estate	5.79***	5.11***	5.44***	4.92***

NOTE:**: p<0.01, ***: p<0.001.

6 Conclusion

Dynamic interaction between various types of individuals in social media is a complex process and the order of network ties is an important aspect of social media network dynamics. We represented various temporal patterns of network formation based on nodal attributes and the order of network ties development and developed NATERGM model for dynamic network analysis. We conducted empirical tests to evaluate the performance of NATERGM and results showed that NATERGM has an enhanced pattern testing capability and potentially better prediction accuracy of network characteristics compared to previous dynamic network models. Compared to existing TERGM-based models, our proposed model can test more complex dynamic patterns resulting from the interaction between network tie formation and nodal attributes, thereby discovering how various nodal attributes are affecting the formation process of a dynamic network. In practice, the proposed model can be used to evaluate the impact of individuals' attributes in the formation process of dynamic social media networks. By examining these attributes, social media designers can understand what factors are critical to the social network evolution and determine what functionalities to add or promote in their platforms.

The contributions of this study are manifold. First, this study provides an extended ERGM-based network model

to examine temporal patterns in dynamic networks. The extended model can examine how nodal attributes of networks affect the order in which network ties develop. Previous models were unable to examine the network dynamics from this perspective. Second, this study provides a list of temporal terms that expands static ERGM terms and dynamic TERGM terms without nodal attributes. The list of temporal terms is designed to be adaptable to any general network. Given a new network, these temporal terms can be used to understand the impact of other nodal attributes beyond the attributes used as examples in this study. Furthermore, this study provides a network prediction framework based on temporal patterns identification, which has been an under-studied area in social network research. In our current model, each temporal pattern only considers one attribute at a time. We plan to extend from this point and consider the interactions of multiple attributes in future research.

ACKNOWLEDGMENT

The research is based upon work supported in part by the National Science Foundation under Grant No. CMMI-1442116 and CMMI-1249210. Thanks to Cathy Larson for her suggestions and comments.

REFERENCES

[1] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Preventing Private Information Inference Attacks on Social Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1849-1862, 2013.

[2] M. Lin, N. R. Prabhala, and S. Viswanathan, "Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending," *Management Science*, vol. 59, no. 1, pp. 17-35, 2013.

[3] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media," *PloS one*, vol. 7, no. 1, p. e29358, 2012.

[4] Z. Shi, H. Rui, and A. B. Whinston, "Content Sharing in a Social Broadcasting Environment: Evidence from Twitter," *MIS Quarterly*, vol. 38, no. 1, 2014.

[5] S. Stieglitz and D.-X. Linh, "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217-248, 2013.

[6] G. C. Kane, M. Alavi, G. J. Labianca, and S. P. Borgatti, "What' Different About Social Media Networks? A Framework and Research Agenda," *MIS Quarterly*, vol. 38, no. 1, pp. 274-304, 2014.

[7] G. Oestreicher-Singer and A. Sundararajan, "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets," *Management Science*, vol. 58, no. 11, pp. 1963-1981, 2012.

[8] P. V. Singh, Y. Tan, and V. Mookerjee, "Network Effects: The Influence of Structural Capital on Open Source Project Success," *MIS Quarterly*, vol. 35, no. 4, 2011.

[9] A. Susarla, J.-H. Oh, and Y. Tan, "Social Networks and the Diffusion of User-Generated Content: Evidence from Youtube," *Information Systems Research*, vol. 23, no. 1, pp. 23-41, 2012.

[10] L. Tang, H. Liu, and J. Zhang, "Identifying Evolving Groups in Dynamic Multimode Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 72-85, 2012.

[11] C.-D. Wang, J.-H. Lai, and P. S. Yu, "Neiwalk: Community Discovery in Dynamic Content-Based Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1734-1748, 2014.

[12] S. Hanneke, W. Fu, and E. P. Xing, "Discrete Temporal Models of Social Networks," *Electronic Journal of Statistics*, vol. 4, pp. 585-605, 2010.

[13] S. Jiang and H. Chen, "A Multi-Theoretical Framework for Hypotheses Testing of Temporal Network Patterns," in *Proceedings of 34th International Conference on Information Systems*, Auckland, New Zealand, 2014.

[14] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network Analysis in the Social Sciences," *Science*, vol. 323, no. 5916, pp. 892-895, 2009.

[15] H. Kwak, C. Lee, H. Park, and S. Moon, "What Is Twitter, a Social Network or a News Media?" in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 591-600.

[16] S. K. Shriver, H. S. Nair, and R. Hofstetter, "Social Ties and User-Generated Content: Evidence from an Online Social Network," *Management Science*, vol. 59, no. 6, pp. 1425-1443, 2013.

[17] S. A. Stewart and S. Abidi, "Applying Social Network Analysis to Understand the Knowledge Sharing Behaviour of Practitioners in a Clinical Online Discussion Forum," *Journal of Medical Internet Research*, vol. 14, no. 6, pp. e170-e170, 2011.

[18] R. H. Atkin, Combinatorial Connectivities in Social Systems: An Application of Simplicial Complex Structures to the Study of Large Organizations: Springer-Birkhäuser: Switzerland, 1977.

[19] A. Traud, P. Mucha, and M. Porter, "Social Structure of Facebook Networks," *Physica A*, vol. 391, no. 16, pp. 4165-4180, 2012.

[20] S. Faraj and S. L. Johnson, "Network Exchange Patterns in Online Communities," *Organization Science*, vol. 22, no. 6, pp. 1464-1480, 2011.

[21] S. P. Borgatti and P. C. Foster, "The Network Paradigm in Organizational Research: A Review and Typology," *Journal of Management*, vol. 29, no. 6, pp. 991-1013, 2003.

[22] L. Qiu, H. Rui, and A. Whinston, "Social Network-Embedded Prediction Markets: The Effects of Information Acquisition and Communication on Predictions," *Decision Support Systems*, vol. 55, no. 4, pp. 978-987, 2013.

[23] M. E. Newman, "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577-8582, 2006.

[24] L. C. Freeman, "Centrality in Social Networks Conceptual Clarification," *Social Networks*, vol. 1, no. 3, pp. 215-239, 1979.

[25] R. S. Burt, Structural Holes: The Social Structure of Competition. Cambridge, Massachusetts: Harvard University Press, 1995.

[26] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1090-1098, 2002.

[27] P. W. Holland and S. Leinhardt. The Statistical Analysis of Local Structure in Social Networks. New York: National Bureau of Economic Research, 1974.

[28] S. Wasserman and P. Pattison, "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs Andp," *Psychometrika*, vol. 61, no. 3, pp. 401-425, 1996.

[29] S. Jiang, Q. Gao, and H. Chen, "The Roles of Sharing, Transfer, and Public Funding in Nanotechnology Knowledge Diffusion Networks," *Journal of the American Society for Information Science and Technology*, 2014.

[30] G. Robins and P. Pattison, "*Interdependencies and Social Processes: Dependence Graphs and Generalized Dependence Structures*," in Models and Methods in Social Network Analysis, P. J. Carrington, J. Scott, and S. Wasserman, Eds. Cambridge: Cambridge University Press, 2005, pp. 192-214.

[31] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An Introduction to Exponential Random Graph P* Models for Social Networks," *Social Networks*, vol. 29, no. 2, pp. 173-191, 2007.

[32] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, "Recent Developments in Exponential Random Graph P* Models for Social Networks," *Social Networks*, vol. 29, no. 2, pp. 192-215, 2007.

[33] Q. Wang, K.-Y. Goh, T. Phan, and S. Cai, "Examining the Timing Effect of Information Diffusion on Social Media Platforms: A Temporal Network Approach," in *Proceedings of the 21th European Conference on Information Systems*, Utrecht, Netherland, 2013.

[34] F. Guo, S. Hanneke, W. Fu, and E. P. Xing, "Recovering Temporally Rewiring Networks: A Model-Based Approach," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007, pp. 321-328.

[35] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating Time-Varying Networks," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 94-123, 2010.

[36] P. N. Krivitsky and M. S. Handcock, "A Separable Model for Dynamic Networks," *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), vol. 76, no. 1, pp. 29-46, 2014.

[37] S. M. Goodreau, D. R. Hunter, C. T. Butts, P. N. Krivitsky, M. S. Handcock, S. B. de-Moll, et al. (2014, 2014). Stergm-Separable Temporal Ergms for Modeling Discrete Relational Dynamics with Statnet.

[38] P. N. Krivitsky, "Modeling of Dynamic Networks Based on Egocentric Data with Durational Information," *Technical Report 2012-01*, Pennsylvania State University Department of Statistics2012.

[39] P. Holme, "Network Dynamics of Ongoing Social Relationships," *Europhysics Letters*, vol. 64, no. 3, pp. 427-433, 2003.

[40] P. Holme and J. Saramäki, "Temporal Networks," *Physics Reports*, vol. 519, no. 3, pp. 97-125, 2012.

[41] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, et al., "Small but Slow World: How Network Topology and Burstiness Slow Down Spreading," *Physical Review E*, vol. 83, no. 2, p. 025102, 2011.

[42] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing Degree Distributions, Clustering, and Homophily in Social Networks with Latent Cluster Random Effects Models," *Social Networks*, vol. 31, no. 3, pp. 204-213, 2009.

[43] P. Sarkar and A. W. Moore, "Dynamic Social Network Analysis Using Latent Space Models," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 31-40, 2005.

[44] P. Sarkar, S. M. Siddiqi, and G. J. Gordon, "A Latent Space Approach to Dynamic Embedding of Co-Occurrence Data," in *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics*, 2007, pp. 420-427.

[45] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, vol. 2, no. 2, pp. 129-233, 2010.

[46] T. A. B. Snijders, "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, vol. 3, no. 2, pp. 1-40, 2002.

[47] R. Zheng, J. Li, H. Chen, and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378-393, 2006.

[48] S. Jiang, H. Chen, J. F. Nunamaker, and D. Zimbra, "Analyzing Firm-Specific Social Media and Market: A Stakeholder-Based Event Analysis Framework," *Decision Support Systems*, vol. 67, no. 1, 2014.

[49] C. Li, "Is Lexical Richness an Essential Criterion in Judging a Piece of Writing?," The University of Hong Kong (Pokfulam, Hong Kong), 1997.

[50] S. Nadarajan, "The Challenges of Getting L2 Learners to Use Academic Words in Their Writings," *Electronic Journal of Foreign Language Teaching*, vol. 8, no. 2, pp. 184-200, 2011.

[51] D. Larsen-Freeman, "Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition," *Applied Linguistics*, vol. 30, no. 4, pp. 579-589, 2009.

[52] K. W. Hunt, Early Blooming and Late Blooming Syntactic Structures. Urbana, IL: National Council of Teachers of English, 1977.

[53] J. Wiebe, T. Wilson, and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165-210, 2005.

[54] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, et al., "Opinionfinder: A System for Subjectivity Analysis," in *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 2005, pp. 34-35.

[55] G. Kossinets and D.J.cWatts, "Empirical analysis of an evolving social network." *Science* vol. 311, no. 5757, pp. 88-90, 2006

Shan Jiang is a doctoral candidate in the Department of Management Information Systems at University of Arizona. He received a B.S. in Management Information Systems from Tsinghua University, China. He is currently working as a research associate in Artificial Intelligence Lab, University of Arizona. His research interests include business intelligence, social media analytics, computational linguistics and social network analysis. His works has appeared in Decision Support Systems, Journal of American Society for Information Science and Technology, Journal of the Association for Information Systems, and International Conference on Information Systems.

Hsinchun Chen received the BS degree from the National Chiao-Tung University in Taiwan, the MBA degree from the State University of New York at Buffalo, and the PhD degree in information systems from New York University. He is a Regent's Professor at the University of Arizona. He has served as a scientific counselor/advisor of the US National Library of Medicine, the Academia Sinica (Taiwan), and the National Library of China (China). He was ranked #8 in publication productivity in information systems (CAIS 2005) and #1 in Digital Library research (IP&M 2005) in two bibliometric studies. His COPLINK system, which has been quoted as a national model for public safety information sharing and analysis, has been adopted in more than 550 law enforcement and intelligence agencies in 20 states. He is a fellow of the IEEE and the AAAS. He received the IEEE Computer Society 2006 Technical Achievement Award.