TTERGM: Social Theory-Driven network simulation

Yifan Huang
Modeling and Simulation Department
University of Central Florida
Orlando, FL, U.S.A

Clayton Barham
Department of Computer Science
University of Central Florida
Orlando, FL, U.S.A

Eric Page
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, FL, U.S.A

Pamela K.Douglas
Modeling and Simulation Department, UCF
Department of Psychiatry and Biobehavioral Medicine, UCLA
Los Angeles, CA, U.S.A

Abstract

Temporal exponential random graph models (TERGM) are powerful statistical models that can be used to infer the temporal pattern of edge formation and elimination in complex networks (e.g., social networks). TERGMs can also be used in a generative capacity to predict longitudinal time series data in these evolving graphs. However, parameter estimation within this framework fails to capture many real-world properties of social networks, including: triadic relationships, small world characteristics, and social learning theories which could be used to constrain the probabilistic estimation of dyadic covariates. Here, we propose triadic temporal exponential random graph models (TTERGM) to fill this void, which includes these hierarchical network relationships within the graph model. We represent social network learning theory as an additional probability distribution that optimizes Markov chains in the graph vector space. The new parameters are then approximated via Monte Carlo maximum likelihood estimation. We show that our TTERGM model achieves improved fidelity and more accurate predictions compared to several benchmark methods on GitHub network data.

1 Introduction

Social networks have been studied for over a century, and graph theory techniques have been used for decades to model relationship patterns amongst individuals and social entities [Wasserman et al., 1994]. Graph network statistics have provided fundamental and theoretical insight into social phenomenal across political, economic and behavioral data. However, historically, these techniques have been applied to study static 'snapshots' of a network, limiting the ability to make large-scale inferences about social network dynamics.

In recent years, online social networks have provided a wealth of open source empirical data, providing new opportunities to test and adjudicate amongst competing theories and quantitative models [Borgatti et al., 2009]. Online social networks are formed by nodes (e.g., individuals) and edges, embedding social relationships (e.g., friends, relatives, colleagues) within a complex graph network. Both

36th Conference on Neural Information Processing Systems (NeurIPS 2022 Temporal Graph Learning Workshop).

network structure and nodal connectivity constraint the flow of information or resources through the network [Kane et al., 2014]. However, online social networks are ever expanding, and their connectivity evolves rapidly over time. Thus, statistically modeling these data for generative and predictive purposes is challenging.

Temporal exponential random graph models (TERGM) have become a core tool for modeling dynamic social networks. Although a variety of approaches including variations of TERGM have been applied to education [Mamas et al., 2020], finance, [Park et al., 2018] and political data [Abrams, 2019], few TERGM variations have attempted to quantify the effect of "influencers" alongside triadic relationships within a dynamic network. Here, we expand the classic TERGM model to support triadic relationships to make predictions on dynamic graphs. We apply this new triadic temporal exponential random graph model (TTERGM) to data derived from Github. Github has been studied extensively within the context of graphical network modeling due to its popularity, open access, and transparency. Thus, these data provide a benchmark for model comparison. GitHub also provides network programming functionality [Borges et al., 2016], offering the opportunity to study "influencers", their affect on network structure, and triadic relationships within a dynamic network.

2 Related Work

Social network models have been applied for many purposes to include: modeling an individual's behavioral patterns to predict future nodal attributes (e.g., connections) over time [McConnell et al., 2018] [McAvoy et al., 2020], modeling interactions and cluster formation within online communities [Fortunato and Hric, 2016] [Xu et al., 2020] [Liu et al., 2018], and modeling how network characteristics (e.g., centrality) influences its users [Qiu et al., 2017]. Overall, these models attempt to characterize the relationship amongst network structure and information diffusion, decision making, and individual behavior. [Jackson et al., 2017].

General classes of network formation methods include: 1) exponential random graph models (ERGMs) [Lusher et al., 2013][Pattison and Wasserman, 1999], meta-networks, and meta-matrices [Carley and Hill, 2001][Krackhardt and Carley, 1998] for multilayer social networks. 2) block modeling [Guimerà and Sales-Pardo, 2009] 3) geographic or characteristic based approaches, [Boucher and Mourifié, 2012][Leung, 2014]; 4) link formation techniques [Christakis et al., 2010][Bramoullé et al., 2012], and 5) subgraph model-based approaches (SUGMs) [Chandrasekhar and Jackson, 2016]. Numerous studies have examined the formation of cascades of network activity, characterizing and predicting network growth [Bakshy et al., 2012][Gruhl et al., 2004][Yang and Counts, 2010]. Typically, spikes of activity occur within a few days of content's introduction into the network. This property forms the backdrop to a line of temporal analyses that focus on the basic rising-and-falling pattern that characterizes the initial occurrence of information burst.

Influence on the Github platform can be quantified by the number of followers, stars, mentions, quotes, and up-votes received from other users. Social network metrics such as centrality indicate how broadly influence extends (e.g. geographic interest) [Weber and Luo, 2014]. Other features include project size, file volume, critical folder, lines of code and calling of basic functions. The popularity rate can be measured by (Total_Stars / project_life). Few studies have examined influence of user-popularity, repo-popularity, and triadic relationships in dynamic graphs.

3 Methodology

3.1 Preliminaries

Exponential random graph models (ERGM) are static. Temporal exponential random graph models (TERGM) are an extension of ERGMs to handle dynamic information in real-world online social networks [Hanneke et al., 2010]. ERGM can be written as in (1). TERGM with Markov assumption can be written as in (2).

$$P(N) = \frac{1}{Z} exp(\sum \gamma(N)) \tag{1}$$

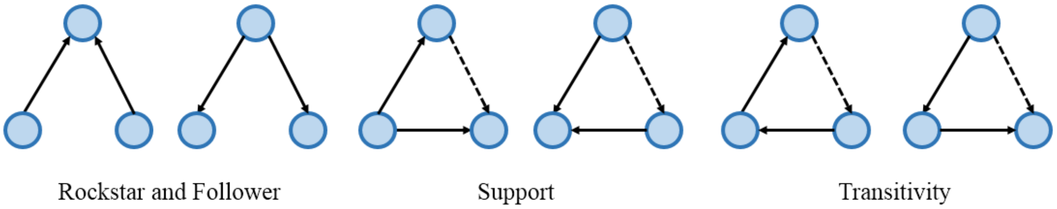

Figure 1: Examples of Triadic relationship in Influencer Networks

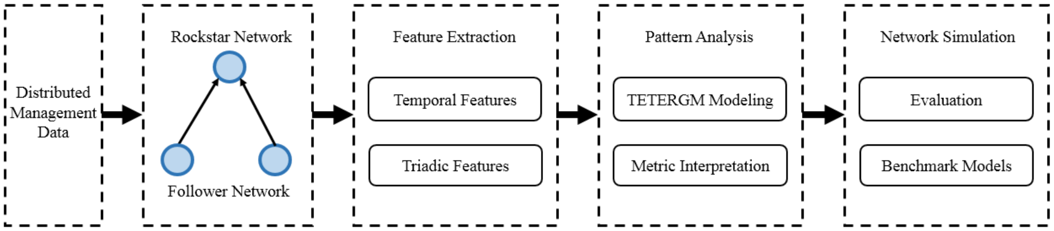

Figure 2: Design of the Triadic Temporal Exponential Random Graph Model

$$P(N^t|N^{t-1}) = \frac{1}{Z_{t-1}} exp(\sum_{ij} \sigma(N^t_{ij}, N^{t-1}_{ij})) \tag{2}$$

$P(N)$ represents the probability of the network N; Z represents the normalization constant that is usually difficult to compute; γ is the vector of network characteristics such as number of edges, triangles, 2-stars, etc. t represents the sequence of network observations; σ is the vector of social-theory driven temporal network characteristics such as homophily, transitivity, reciprocity, etc. t represents the sequence of network observations; Compared to ERGMs, TERGMs are able to model the distribution on time series data (either embedded in the network or in separate timesteps), hence certain temporal patterns can be captured and reflected in the parameter values for phenomenon interpretation. Examples of these dynamic patterns of triadic effects in influencer networks are shown in Figure 1. TERGM provides significant advantages over ERGMs, since certain static patterns can be enriched in higher dimensions when the sequence order is considered. TERGMs are also capable of modeling observed friendship networks with bootstrap methods estimated by maximum pseudolikelihood [Leifeld et al., 2018], or networks of infectious disease transmission using statistical methods in network analysis [Jenness et al., 2018]. The flexibility of TERGMs make it possible to adapt to a variety of input data types, such as cross-sectional or longitude data [Henry et al., 2016][Block et al., 2018].

3.2 Triadic Temporal Exponential Random Graph Model

TERGM models estimated within the markov chain assumption are typically incapable of generating and reproducing realistic dynamics observed in real-world online social networks. We hypothesized that increasing the model's capacity to describe triadic network properties would reduce the error between the model and empirical observations. We propose TTERGM here to sequentially predict network probabilities by integrating the dynamics between influencers and followers. TTERGM was run on a computer with 12900K CPU, 1080TI and 128GB RAM. Figure 2 shows the framework of TTERGM that has five major components - data collection module, network processing module, feature extraction module, pattern analysis module, and a generative network module.

Influencer-follower networks N_t were identified and constructed by connecting users by events in Table 1. In the subsequent module, temporal features and triadic features listed in Table 2 were extracted from the influencer-follower networks. Network characteristics were then estimated using the Markov chain Monte Carlo (MCMC) method. In the pattern analysis module, the TTERGM model was applied to model the data. The general form of TTERGM can be written as in Equation (3). $P(N)$ represents the probability of a given network architechture, N; Z represents the normalization constant as done in classic TERGM models; γ is a vector of network characteristics (e.g, number of edges, triangles, 2-stars), t represents the temporal sequence of network observations, and σ is the vector of social-theory driven temporal network characteristics such as homophily, transitivity, reciprocity, etc. To fit the TTERGM model to the data, algorithm 1 is used to initialize and traverse each node in the network to construct the generative influencer-follower networks. Finally, we used these observed features to simulate real-world influencer-follower networks and compared the predictive of TTERGM performance with the classic TERGM model and the Block Model using left-out validation data.

$$P(N^t|N^{t-1}) = \frac{1}{Z_{t-1}}exp(\sum_{ij}\sigma(N_{ij}^t, N_{ij}^{t-1})) + ... + \frac{1}{Z_1}exp(\sum_{ij}\sigma(N_{ij}^2, N_{ij}^1)) \tag{3}$$

Algorithm 1: Triadic Temporal Exponential Random Graph for Network Generation

```
Input: Online social network $N$
Initialization $G = G_{t_0}$ repeat
    while $t \neq t_n$ do
        read current node $n_i$;
        if $n_i$ is not traversed then
            for each link $l_{ij}$ in graph $G_{t_0}$ do
                calculate the network statistics of each link and
                fill the values to the feature vector $N_{ij}$
                $Prob(l_{ij}) = \sum_{p=0,q=0}^{pq} Prob(l_{ij}|l_{pq})$
        else
            read the next node $n_{i+1}$;
        $t = t + 1$;
    Calculate $P(N^t) = \frac{1}{Z}exp(\sum_{ij}\sigma(N_{ij}^t, N_{ij}^{t-1}, ..., N_{ij}^{t_0})$ where
    $Z = MCMC(exp(\sum_{ij}u(N_{ij}^t, N_{ij}^{t-1}, ..., N_{ij}^{t_0}))$ where
    $Z$ is the normalization constant, $u$ is the network statistics, $MCMC$ is the
     Markov Chain Monte Carlo method
until k iterations;
```

4 Experiments and Results

The GitHub dataset is from 1/1/2015 to 8/30/2017 including 2 million users and 13 million projects. We selected the 100 most popular repositories because we aimed to characterize top GitHub repositories, and the impact of influencers on the popularity of repositories using the TTERGM model. The number of repositories is a hyperparameter which can be adjusted, depending on the goal of the model. We selected the top 10 users, in terms of number of followers, as the influencers in this study. This threshold was chosen because the number of followers drops off sharply after that point, but can be chosen arbitrarily for different datasets. The data was acquired using the API from GitHub [Gousios and Spinellis, 2012]. The API can be used to stream GitHub repository interactions with customized formats. Optionally, meta data from user relation events can be retrieved as well. The dataset contains 14 types of events which are listed in Table 1.

We implemented the TTERGM technique to model the dynamics between repository popularity and networks. The 3 most followed anonymized influencers had 52,722, 30,161, and 25,827 followers respectively. Each of the top 10 most popular influencer has at least 14 thousand followers. We divided the 15 types of events into 2 categories - participative events and contributive events to highlight the social characteristics. Participative events demonstrate a user's engagement to a project repository. Contributive events indicate the cooperation among developers in a software repository. Followers tend to have more participative events than influencers, while influencers generally have more contributive events than followers.

Table 1: Event Categories and Descriptions

Event Category	Event Type	Description
Receptive Events	WatchEvent	When someone stars a repository
	PullRequestReviewCommentEvent	When comment on a pull request's unified changes
	IssueCommentEvent	When an issue comment is created or edited
	MemberEvent	When a user is invited or removed as a collaborator to a repository
	IssuesEvent	When an issue is created or edited
	GollumEvent	When a wiki page is created or edited
Contributive Events	WatchEvent	When someone stars a repository
	ForkEvent	When an user forks a repository
	ReleaseEvent	When a release is published or edited
	PublicEvent	When a private repository is made public
	PullRequestEvent	When a pull request is created or edited
	PushEvent	When a push is happened to a repository branch
	DeleteEvent	When a branch or tag is deleted
	CommitCommentEvent	When is commit comment is created
	CreateEvent	When a branch or tag is created

Table 2: Features Extracted from Influencer Network

Measurement Type	Measurement Name	Description
Network connection	# direct links	Number of direct links influencers have with their followers
	# indirect length-2 links	Number of indirect links of length 2 from influencers to their followers
	# indirect length-3 links	Number of indirect links of length 3 from influencers to their followers
	# triangles	Influencer's activity on a repository triggered his/her follower's activity on the same repository, then a triangle forms
Network topology	Average shortest path	Average shortest path length of pairs of nodes in the network
	Assortativity	Pearson correlation coefficient of degree between pairs of linked nodes
	# of connected components	Number of connected components in the network
	Average Clustering Coefficient	Ratio of number of triangles over maximum possible number of triangles
	# nodes	Number of nodes in the network
	# edges	Number of edges in the network

To evaluate the performance of the TTERGM, we compared the simulation with two benchmark models - TERGM and Block model. Each model is evaluated using a set of metrics (average degree of incoming edges and average degree of outgoing edges) to see how well the predicted network distribution matched the real-world network characteristics. The evaluation result (average of 30 runs) was shown in Table 4. The Block model and TERGM perform similarly in month 2017-07 and

Table 3: Number of Followers of the Top 10 Influencers

Rank	Anonymized Influencer Id	Number of Followers
1	lBMOoXAjxIN_Dc3alQNLZQ	52722
2	BhQS5KA8AvmQJXbsVeusdw	30161
3	s0jAeLRt2onrivaUCqdJrg	25827
4	QFB1aZ8GXkNYHyfWe7aEeA	24604
5	jAGnWUFUmnBc9ydeQbIfDQ	24510
6	hXalEIoEWnEbCSfiQI1LNA	23076
7	eUnkVgArKJiNOBhb0w53_Q	18522
8	VRyyOPSJUCS5jRlDtwjefA	15755
9	wNDkYd6NACSuvLCnxog23w	15396
10	wHfAzUFXU8D186qTl9c54w	14928

Table 4: Block, TERGM, and TTERGM models were used to generate predicted distributions for network characteristics on out-of-sample temporal observations

Network Model	2017-07 in-deg	2017-07 out-deg	2017-08 in-deg	2017-08 out-deg
Block Model	4.39	5.08	5.32	4.56
TERGM	4.65	4.78	5.13	4.35
TTERGM	3.42*	4.15*	4.25*	3.25*

Note: $*$ indicates p-value < 0.05 comparing to Block Model and TERGM respectively

2017-08. Comparing to TERGM, TTERGM has 26.45% less errors in 2017-07 for incoming degree of errors, 13.17% less errors in 2017-07 for outgoing degree of errors, 17.15% less errors in 2017-08 incoming degree of errors, and 25.28% less errors in 2017-08 for outgoing degree of errors. We believe the consistent improvement stems from the extra computation from TTERGM in the markov chain.

5 Conclusion

We implemented a social-theory driven temporal exponential random graph model to infer the temporal pattern of edge formation and elimination in complex networks (e.g., social networks), and examine the effect of influencers and triadic relatinships on predicting future network dynamics. When popular repositories are formed or influencers act, the structure of the social network alters, affecting network metrics. The TTERGM technique build upon previous statistical models by incorporating information flow across hierarchical configuration features. We represent social network learning theory as an additional probability distribution that optimizes Markov chains in the graph vector space. The new parameters are then approximated via Monte Carlo maximum likelihood estimation. The TTERGM model is capable of reproducing the dynamics observed empirically in large-scale social network data, and produced more accurate predictions on left-out data compared to the classic TERGM and block models. However, the TTERGM model imposes additional computational burden during parameter estimation, which may hinder its ability to scale to larger datasets. Future work may include expanding this approach to model the influence of more "distant" users in the network, or those that do not directly follow an influencer.

References

A. Abrams. Here's what we know so far about russia's 2016 meddling. *Time, https://time.com/5565991/russia-influence-2016-election/*, 2019.

E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

P. Block, J. Koskinen, J. Hollway, C. Steglich, and C. Stadtfeld. Change we can believe in: Comparing longitudinal network models on consistency, interpretability and predictive power. *Social Networks*, 52:180–191, 2018.

S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.

H. Borges, A. Hora, and M. T. Valente. Understanding the factors that impact the popularity of github repositories. In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 334–344. IEEE, 2016.

V. Boucher and I. Mourifié. My friend far far away: Asymptotic properties of pairwise stable networks. *Available at SSRN*, 2170803(1.1):6, 2012.

Y. Bramoullé, S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers. Homophily and long-run integration in social networks. *Journal of Economic Theory*, 147(5):1754–1786, 2012.

K. M. Carley and V. Hill. Structural change and learning within organizations. *Dynamics of organizations: Computational modeling and organizational theories*, pages 63–92, 2001.

A. G. Chandrasekhar and M. O. Jackson. A network formation model based on subgraphs. *Available at SSRN 2660381*, 2016.

N. A. Christakis, J. H. Fowler, G. W. Imbens, and K. Kalyanaraman. An empirical model for strategic network formation. Technical report, National Bureau of Economic Research, 2010.

S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics reports*, 659: 1–44, 2016.

G. Gousios and D. Spinellis. Ghtorrent: Github's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, pages 12–21. IEEE, 2012.

D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.

R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

S. Hanneke, W. Fu, E. P. Xing, et al. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

T. Henry, S. B. Gesell, and E. H. Ip. Analyzing heterogeneity in the effects of physical activity in children on social network structure and peer selection dynamics. *Network Science*, 4(3):336–363, 2016.

M. O. Jackson, B. W. Rogers, and Y. Zenou. The economic consequences of social-network structure. *Journal of Economic Literature*, 55(1):49–95, 2017.

S. M. Jenness, S. M. Goodreau, and M. Morris. Epimodel: an r package for mathematical modeling of infectious disease over networks. *Journal of statistical software*, 84, 2018.

G. C. Kane, M. Alavi, G. Labianca, and S. P. Borgatti. What's different about social media networks? a framework and research agenda. *MIS quarterly*, 38(1):275–304, 2014.

D. Krackhardt and K. M. Carley. *PCANS model of structure in organizations*. Carnegie Mellon University, Institute for Complex Engineered Systems . . . , 1998.

P. Leifeld, S. J. Cranmer, and B. A. Desmarais. Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software*, 83(6), 2018.

M. Leung. A random-field approach to inference in large models of network formation. *Available at SSRN*, 2014.

Y. Liu, L. Li, H. Wang, C. Sun, X. Chen, J. He, and Y. Jiang. The competition of homophily and popularity in growing and evolving social networks. *Scientific reports*, 8(1):1–15, 2018.

D. Lusher, J. Koskinen, and G. Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2013.

C. Mamas, P. Bjorklund Jr, A. J. Daly, and S. Moukarzel. Friendship and support networks among students with disabilities in middle school. *International Journal of Educational Research*, 103: 101608, 2020.

A. McAvoy, B. Allen, and M. A. Nowak. Social goods dilemmas in heterogeneous societies. *Nature Human Behaviour*, pages 1–13, 2020.

E. McConnell, B. Néray, B. Hogan, A. Korpak, A. Clifford, and M. Birkett. "everybody puts their whole life on facebook": Identity management and the online social networks of lgbtq youth. *International journal of environmental research and public health*, 15(6):1078, 2018.

H. Park, M. A. Bellamy, and R. C. Basole. Structural anatomy and evolution of supply chain alliance networks: a multi-method approach. *Journal of Operations Management*, 63:79–96, 2018.

P. Pattison and S. Wasserman. Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193, 1999.

L. Qiu, H. Kenneth Cheng, and J. Pu. Hidden profiles in corporate prediction markets: The impact of public information precision and social interactions. *MIS Quarterly*, 41(4), 2017.

S. Wasserman, K. Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

S. Weber and J. Luo. What makes an open source code popular on git hub? In *2014 IEEE International Conference on Data Mining Workshop*, pages 851–855. IEEE, 2014.

X. Xu, H. Qian, C. Ge, and Z. Lin. Industry classification with online resume big data: A design science approach. *Information & Management*, 57(5):103182, 2020.

J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.