J Nonlinear Sci
https://doi.org/10.1007/s00332-018-9465-y

Journal of Nonlinear Science

The Role of Graphlets in Viral Processes on Networks

Samira Khorshidi[1] · **Mohammad Al Hasan**[1] · **George Mohler**[1] · **Martin B. Short**[2]

Received: 7 February 2018 / Accepted: 3 May 2018

Abstract Predicting the evolution of viral processes on networks is an important problem with applications arising in biology, the social sciences, and the study of the Internet. In existing works, mean-field analysis based upon degree distribution is used for the prediction of viral spreading across networks of different types. However, it has been shown that degree distribution alone fails to predict the behavior of viruses on some real-world networks and recent attempts have been made to use assortativity to address this shortcoming. In this paper, we show that adding assortativity does not fully explain the variance in the spread of viruses for a number of real-world networks. We propose using the graphlet frequency distribution in combination with assortativity to explain variations in the evolution of viral processes across networks with identical degree distribution. Using a data-driven approach by coupling predictive modeling with viral process simulation on real-world networks, we show that simple regression models based on graphlet frequency distribution can explain over 95% of the variance in virality on networks with the same degree distribution but different

Communicated by Mason A. Porter and Andrea L. Bertozzi.

✉ Martin B. Short
mbshort@math.gatech.edu

Samira Khorshidi
sakhors@iu.edu

Mohammad Al Hasan
alhasan@cs.iupui.edu

George Mohler
gmohler@iupui.edu

[1] Computer and Information Science, Indiana University - Purdue University Indianapolis, Indianapolis, USA

[2] School of Mathematics, Georgia Institute of Technology, Atlanta, USA

Published online: 26 May 2018

network topologies. Our results not only highlight the importance of graphlets but also identify a small collection of graphlets which may have the highest influence over the viral processes on a network.

Keywords Graphlets · Viral processes · Hawkes process · SIS model

Mathematics Subject Classification 68R10 · 91D30 · 60G99

1 Introduction

A variety of dynamic phenomena, including Youtube video views (Crane and Sornette 2008), Tweet resharing (Zhao et al. 2015), viral marketing campaigns (Leskovec et al. 2007), the spread of computer viruses on the Internet (Berger et al. 2005), and gang retaliation (Short et al. 2014) can be explained as evolving viral processes on networks. As such, the study of the evolution of viral processes on networks has attracted considerable attention in recent years. It is now well known that for a connected network, the largest eigenvalue of its adjacency matrix is a good metric for predicting the viral process in that network (Ganesh et al. 2005; Chakrabarti et al. 2008a; Yang et al. 2015). The largest eigenvalue can be roughly estimated by the average degree of the network (Lovasz 2007), but the complete degree distribution of the network is more expressive than the point estimate of average degree and has therefore also been considered for predicting these viral processes. A common approach along these lines is to employ a mean-field analysis where independence assumptions on the nodes are used (Dietz 1980; Anderson et al. 1992; Dezső and Barabási 2002; Callaway et al. 2000; Chakrabarti et al. 2008b). More recently it has been shown that in some cases, including real-world networks, these mean-field assumptions fail to predict the viral spreading (Givan et al. 2011). Consequently, assortativity has been proposed for addressing the limitations of the degree-based, mean-field analyses (Van Mieghem et al. 2010; Jalan and Yadav 2015), and through degree-preserving network rewiring procedures, it has been shown that the spectral radius of a graph can exhibit great fluctuations across networks with the same degree distribution and that assortativity can be used to explain this variation. While there is a clear correlation between assortativity and the dynamics of viral processes on graphs, the fact remains that, if we control for both assortativity and degree distribution, viral diffusions on networks with differing topologies can still exhibit greatly different behaviors.

We illustrate this observation with the following example. In Fig. 1, we display results for a Hawkes process (Crane and Sornette 2008) simulation on 2500 networks with identical degree distribution sampled via degree-preserving rewiring (Qu et al. 2015) from the Karate network (Rossi and Ahmed 2015). In the Hawkes branching process model, an initial event occurs at a randomly chosen node and then subsequent generations of events occur at neighboring nodes of previous events with a fixed probability. In Fig. 1, we plot assortativity versus the expected total number of events (at time infinity) of the Hawkes process for each of the simulated networks. While assortativity partially explains the behavior of the process, for the two highlighted networks with identical degree distribution and similar assortativity, the expected total

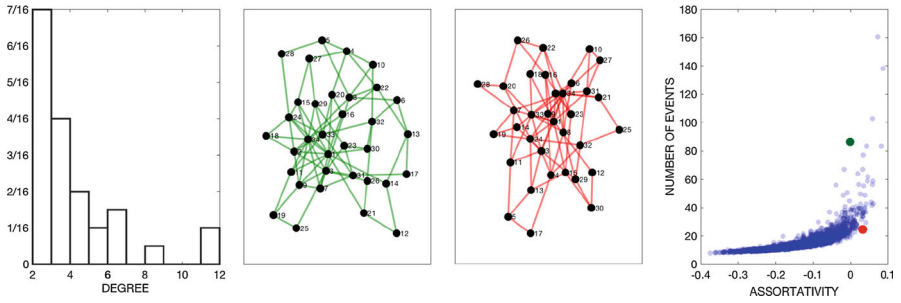

Fig. 1 Hawkes process simulation on 2500 rewired Karate networks. Degree distribution (far left) is fixed for all of the 2500 networks. The expected number of events in a cascade versus assortativity (far right). Two example networks (middle) corresponding to large differences in virality despite similar assortativity and identical degree distribution

number of events in the process differs by a factor of 4. So, we need to extend our analysis beyond degree distribution and assortativity for better understanding of the dynamics of a viral process over a network.

In this paper, we propose using the frequency distribution of graphlets (see Fig. 3 for a preview) to explain the variation in viral processes observed in Fig. 1. Specifically, we show that graphlet frequencies are good predictors for explaining the variation of the evolution of viral processes over a collection of networks for which the degree distribution is kept fixed. Our results not only highlight the importance of graphlets but also identify a small collection of graphlets that may have the highest influence over the viral processes on a network.

The rest of the document is organized as follows. In Sect. 2, we discuss some background materials, including graphlets and viral process models. In Sect. 3, we discuss the methodologies of our proposed analysis. In Sect. 4 we present our results on five real-world networks illustrating the role of graphlets in viral processes. In the final section, we discuss the implications of these results and directions for future research.

2 Background

In this paper, we will be making several different measurements reflecting the topology of networks—from degree distribution to assortativity to graphlet distribution—as well as simulating two different viral processes—the Hawkes process and Susceptible–Infected–Susceptible model—on networks. We provide some background materials on these topics in this section.

2.1 Degree Distribution

The *degree* of a node is the number of connections of that node to other nodes in the network. The degree distribution is a discrete probability distribution of degrees over the nodes in the network. Directed networks have two different degree distributions, the in-degree and the out-degree distributions. In this study, we restrict our attention to undirected networks for which only one degree distribution is defined.

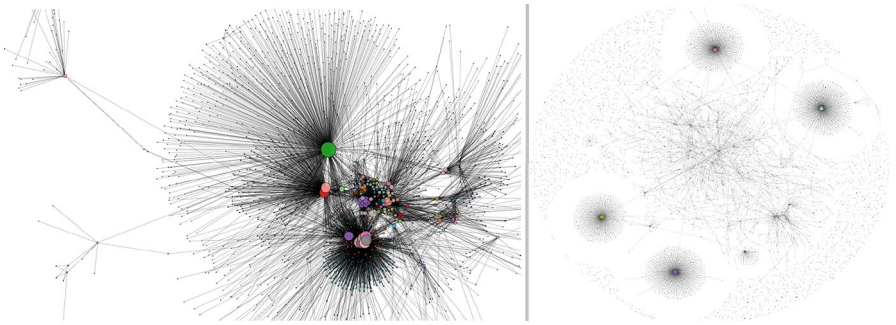

Fig. 2 The left figure shows a graph with positive assortativity where nodes with high degree are connected to other nodes with high degree (assortativity = 5.4). The right graph is an example of a disassortative network (assortativity = −0.88), where the hubs connect to low degree nodes

2.2 Assortativity

Assortativity, or assortative mixing, is defined by the tendency of a network's nodes to be connected to other nodes that are similar in some way. While there are several different mathematical definitions, we will refer to assortativity as the Pearson correlation coefficient of degrees at either ends of a network edge (Newman 2002). In this case the assortativity is given by the formula (Newman 2002),

$$A = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}, \tag{1}$$

where the edges in a network are indexed by $i = 1, \ldots, M$ and j_i, k_i are the degrees of the nodes at the ends of edge i. In Fig. 2, we provide examples of two social networks from the Network Repository (Rossi and Ahmed 2015) with different assortativity, one positive and one negative.

2.3 Graphlets

Graphlets can be defined as small, connected,[1] non-isomorphic, induced subgraphs of a large network. In this study, we work with all possible graphlets having $k \in \{3, 4, 5\}$ vertices. If the graphlet edges are undirected, there are 29 such graphlets as shown in Fig. 3. We refer to a graphlet with k vertices as a k-Graphlet; note that a 1-Graphlet is simply a vertex and a 2-Graphlet is simply an edge. The frequency of a graphlet g_i in a graph G is the total number of distinct embeddings of that graphlet g_i in the graph G. The Graphlet frequency distribution (GFD) is the normalized frequency of the graphlets. To obtain GFD, we normalize the graphlet count vector (a vector of 29

[1] Disconnected graphlets have also been considered in several graphlet-based works, but in this work we only consider connected graphlets because connectedness is essential for the evolution of a viral process on a network.

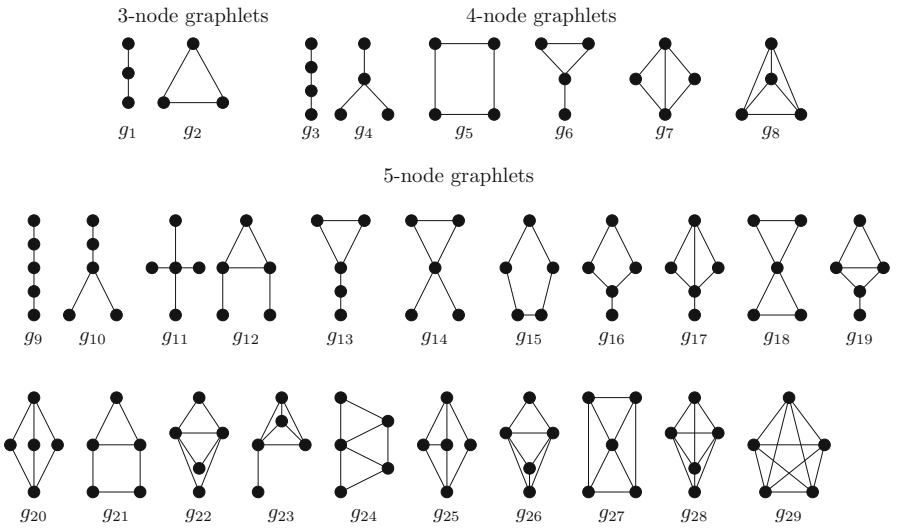

Fig. 3 Undirected graphlets with 3, 4, and 5 vertices

integers which represents the count of each of the graphlets) so that the $L1$-norm of the vector is 1; thus, the values in the vector represent a discrete probability distribution. As real-life networks are generally sparse, frequencies of larger-sized graphlets shrink exponentially, so in GFD, we typically use a logarithm scale for comparing various frequencies so that the number of occurrences of larger-sized graphlets against that of the smaller-sized ones are scaled appropriately.

The frequency distribution of graphlets in a network captures the local topology around the vertices of the network and it has a number of uses in network analysis and prediction. For example, the frequencies of various graphlets can be used for building a global fingerprint for a network such that the fingerprints of different networks arising from a real-life domain are almost identical (Rahman et al. 2014a). Transition of graphlets over temporal snapshots of a network has also been shown to improve link prediction models for dynamic networks (Rahman 2016).

A key challenge for finding graphlet frequency distribution is the high computational cost of graphlet enumeration or counting. However, in recent years, several efficient algorithms have become available that generate exact graphlet counting (Rahman et al. 2012, 2014a). Yet, for very large networks exact counting of graphlets is not feasible. So, there exist methods [such as, GUISE and its variants (Rahman and Al Hasan 2014; Bhuiyan et al. 2012; Rahman et al. 2014b)], which can generate approximate graphlet frequency distributions through uniform sampling of graphlets. Sampling-based methods provide very good approximations of graphlet frequency distributions and can scale to graphs with millions of vertices. In this work, we use GUISE algorithm for computing the graphlet frequency distributions of a network.

2.4 Hawkes Process Model

The Hawkes process is a specific kind of self-exciting point process whereby discrete events occur according to a stochastic intensity $\lambda(t)$ that increases in response to events

themselves. For a Hawkes process on a network G, given the event observation dyads (t_i, v_i), where t_i is a time value for an event and v_i is a vertex on which the event occurred, the conditional intensity (rate) of events at node v, $\lambda_v(t)$, is given by the equation

$$\lambda_v(t) = \mu + \sum_{\substack{t > t_i \\ v_i \in N(v)}} \theta f(t - t_i). \tag{2}$$

In Eq. 2, $N(v)$ is the set of nodes that are neighbor of v on G (note that $v \in N(v)$ for this process) so that the intensity is a superposition of a Poisson background intensity μ and Poisson intensities $\theta f(t - t_i)$ centered at previous event times t_i such that v_i is a neighbor of v on the graph G. The triggering kernel $f(t)$ is a probability density defined on $[0, \infty)$ and, when the model is interpreted as a branching process, the productivity parameter θ determines the expected number of direct offspring events at node v triggered by an event at a node $v_i \in N(v)$.

2.5 Susceptible–Infected–Susceptible Model

The second model we consider is based on the well-known Susceptible–Infected–Susceptible (SIS) model in epidemiology. In this model, each node of the network can be in one of two states at any given time—susceptible or infected. Let $S(t)$ be the set of all susceptible nodes at time t, $I(t)$ be the set of all infected nodes at time t, $\lambda_v(t)$ be the rate at which a susceptible node $v \in S(t)$ becomes infected, and μ_u be the rate at which an infected node $u \in I(t)$ becomes susceptible. The model is then

$$\lambda_v(t) = \sum_{q \in N(v) \cup I(t)} \theta \tag{3}$$

$$\mu_u = 1. \tag{4}$$

So, a susceptible node v switches to being infected via a Poisson process with time varying rate equal to a parameter θ times the number of neighbors of v that are infected at that time, and an infected node u switches to being susceptible via a homogeneous Poisson process with rate 1 (without loss of generality). In terms of virality, one may be interested in a scenario in which all nodes are initially susceptible except for a potentially small number of infected, then tracking how many further infections occur as a result of these initial infections. The result will clearly depend on which nodes are initially infected, the adjacency matrix of the network A, and the parameter θ.

3 Methodologies

Our primary objective is to show that graphlet frequency distribution, in addition to assortativity, is a good predictor for the virality in a network when degree distribution is controlled for. Many earlier works use analytic approaches for finding the influence of network topology or network-based metrics on the viral process on a network (Ganesh et al. 2005; Chakrabarti et al. 2008a). However, such methods are very cumbersome

for graphlets, as graphlets are combinatorially complex objects and their influence over the dynamic process is difficult to represent by a simple model that can be solved analytically. So, in this work we forgo a mathematical analysis in favor of a data science approach to the problem. Our overall strategy is to use simulation and empirical measurement to explore the connection between graphlet distribution and the evolution of viral processes on networks.

For this purpose, we use several real-world networks as input to a simulation model. Given a particular network and model for a viral process, we perform the following steps:

1. Generate M synthetic networks through rewiring (explained below) with identical degree distributions to the original real-world network.
2. For each generated network, compute the assortativity and graphlet frequency distribution.
3. For each generated network, compute the expected number of events, $E[N_\infty]$, of the viral process of interest running on the network.
4. Regress $E[N_\infty]$ against the assortativity and/or graphlet frequency distribution to assess the role that these measures play beyond degree distribution.

Below we provide more details of the above steps.

3.1 Generating Simulation Graphs

We have shown in Sect. 1, degree distribution and assortativity are not adequate for explaining the evolution of a viral process in a network—which motivate us to find the influence of graphlet frequency distribution in a viral process. However, degree distribution does provide a partial explanation of a viral process. To nullify the influence of degree distribution in our analysis, we use degree distribution as a control variable, i.e., we generate a collection of synthetic networks for which degree distribution is a constant.

Generating networks with a given degree distribution are a well-studied problem, specifically for the task of network motif discovery (Saha and Al Hasan 2015). There are two well-known approaches for solving this problem, (1) edge swapping (Mihail and Zegura 2003; Maslov and Sneppen 2002) and (2) stub-matching. Edge swapping starts from a given graph and makes local modification on the given graph to generate another graph having the same degree sequence. One edge swapping approach that preserves the degree sequence is the following. First, select two edges uniformly at random from the graph G; for example, suppose these are $e_1 = (a, b)$ and $e_2 = (c, d)$. Then, replace these two edges by two new edges where the second vertices are swapped between the original two edges, assuming those new edges are not already present in G; in our example, these would be the two new edges $e_3 = (a, d)$ and $e_4 = (c, b)$. If the edge e_3 or e_4 (or both) already exists in G, this proposed swap is rejected and the process is repeated with a new pair of randomly chosen edges e_1 and e_2. It is easy to see that the degree of each vertex remains invariant under a successful edge swap process. The edge swapping can be continued and a sequence of graphs can be generated, such that all of these graphs have an identical degree distribution. If we consider the sequence of graphs as a Markov chain, then the stationary distribution of the Markov

chain is a uniform distribution over the graphs having identical degree distribution. For stub-matching, the configuration model is very popular (Newman 2003). In this method, the algorithm creates as many stubs (dangling half-edges) for each vertex as its degree. Then edges are created by choosing pairs of vertices randomly and connecting their stubs. This approach may create parallel edges, which are dealt with by restarting the process; for large graph the restarting may become very costly.

In this work, we use the edge swapping method, as it is easy to implement. By choosing a sufficiently large number of steps for the Markov chain, we generate graphs which are sufficiently different from each other with widely different graphlet frequency distributions.

3.2 Preparing Topology and Virality Data for Regression

For our study, a collection of graphs with identical degree distribution is a regression dataset in which each graph is an instance. For each graph, we compute graphlet degree distribution and assortativity, which become the explanatory variables for our regression. Below we discuss how we compute these values for a given graph.

Computing graphlet frequency distribution by counting each of the graphlets in a graph is a costly task as the number of graphlet embeddings grows exponentially with the size of the graphlets. In fact, if both connected and disconnected graphlets are considered, the number of k-graphlets on a graph with K vertices is equal to $O(\binom{K}{k})$. In this work, we consider graphlets up to size 5, for which a brute-force graphlet enumeration complexity is equal to $(O(K^5)$, which is not scalable for many real-life networks. An alternative to counting is uniform sampling of graphlet embeddings, which is sufficient to obtain a graphlet frequency distribution. In an earlier work (Rahman et al. 2012), we have shown how graphlets can be sampled under uniform distribution by using a Monte Carlo Markov chain (MCMC) sampling algorithm. Specifically, we have proposed a method named GUISE, which performs a random walk over the graphlet embeddings by following a double-stochastic transition matrix; the stationary distribution of the Markov chain is a uniform distribution over the graphlet embeddings. By counting the type of graphlets that are traversed in this random walk and then normalizing the vector as described above, GUISE returns a graphlet frequency distribution vector. In this work, we use GUISE algorithm for computing the graphlet frequency distribution. It returns a 29-size vector, in which each component $d(g_i)$ represents the normalized frequency of the graphlet g_i as illustrated in Fig. 3. We also compute the assortativity of the network using Eq. 1. The components of the graphlet frequency distribution vector and assortativity (a 30-size vector) become the co-variates of our regression analysis.

The target value of our regression is the expected number of events—excited offspring events in the case of the Hawkes process and secondary infections in the case of the SIS process—that are spawned from a single initiating event placed randomly within the network. The way this target value is computed depends on the viral process used. For the Hawkes process (Crane and Sornette 2008; Zhao et al. 2015; Short et al. 2014), we consider a simplified model where

1. a node is chosen uniformly at random
2. an initial event at time $t_1 = 0$ occurs at the chosen node
3. the Hawkes process with $\mu = 0$ is simulated and the total number of events, N_∞, at time infinity is observed

In the case of this simplified model, the expected number of total events is given by,

$$E[N_\infty] = \frac{1}{K} 1^T \cdot \left(\sum_{j=1}^{\infty} (\theta A)^j \cdot 1 \right), \tag{5}$$

where K is the number of nodes in G, 1 is a column vector of ones, and A is the adjacency matrix of G (A is symmetric because the graph G is undirected). The expected total number of events $E[N_\infty]$ will be finite up to a critical threshold value of the productivity parameter, θ_c.

For the SIS model, we approximate the continuous time version described in Sect. 2 above with a discrete time version that allows us to more easily count the number of secondary infections arising from a single initially infected node, and which greatly simplifies the simulations. Here, time is discretized into units of step 1, and we define a vector $I(t)$ such that $I_u(t) = 1$ if node u is infected at time t and 0 if node u is susceptible at time t. Initially, $I_u(0)$ is zero for all u except for a single node chosen uniformly randomly. Then the model proceeds via iterations of the following steps:

1. nodes v susceptible at time t become infected at time $t + 1$ with probability $1 - e^{-\lambda_v(t)}$, where $\lambda_v(t) = \theta(AI(t))_v$
2. all nodes infected at time t become susceptible at time $t + 1$
3. at each timestep $t > 0$, the product $1^T \cdot I(t)$ is equal to the new number of infections
4. the total new infections N_∞ are observed as time goes to infinity

For this simplified model, an approximation of the expected number of total new infections can be found via

$$E[N_\infty] = \frac{1}{K} 1^T \cdot \sum_{k=1}^{K} \sum_{t=1}^{\infty} I(t; k) \tag{6}$$

$$I_v(0; k) = \delta_{v,k} \tag{7}$$

$$I_v(t + 1; k) = (1 - I_v(t; k)) \left[1 - e^{-\theta(AI(t;k))_v} \right]. \tag{8}$$

As in the Hawkes model, the value of $E[N_\infty]$ is expected to be finite up until some critical value of θ, below which the infection is expected to disappear at some finite time (is at most epidemic), and above which the infection in expectation never leaves the network (is endemic).

3.3 Regression Model for Predicting Virality

For each real-world network, we simulate $M = 500$ rewired networks holding the degree distribution fixed. Next, for each simulated network, we compute the assor-

tativity, the graphlet frequency distribution, the largest eigenvalue $\lambda_{\max}$, as well as $E[N_\infty]$ for the Hawkes and SIS models. When calculating $E[N_\infty]$ for the Hawkes and SIS models, the value of θ is chosen to be a constant multiple of the largest eigenvalue of the original network. To explain the observed variation in $E[N_\infty]$ across the simulated networks, we run a regressions of the form

$$\log(E[N_\infty]) = b + c_0 A + \sum_{i=1}^{29} c_i \log(d(g_i)) + \epsilon \tag{9}$$

where b is the intercept, A is the assortativity, $d(g_i)$ is the frequency distribution value of graphlet g_i, and c_i are coefficients of a linear regression where the model errors ϵ are assumed to be normal. Note that, although our regression is simply a linear fit, the underlying relationship between virality and the various graph topology measures is proposed to be nonlinear due to the logarithms present in Eq. 9. This nonlinearity is certainly plausible at least for the assortativity, given the plotted relationship in Fig. 1. Further, as discussed above, it is natural to consider the logarithm of the graphlet frequency distribution, which is why we do so here. We estimate the model on 70% of the 500 simulated networks and then evaluate the R^2 and mean square error (MSE) on the remaining 30% test data.

4 Data and Results

We consider four real-world networks obtained from the Network Repository (Rossi and Ahmed 2015). The networks include (1) a retweet network where the nodes are twitter users and edges are retweets (collected from various social and political hashtags) Rossi et al. (2012); (2) a karate network where the dataset contains social ties among the members of a university karate club collected by Wayne Zachary in 1977 Zachary (1977); (3) a social network of bottlenose dolphins where the dataset contains a list of all the links (a link represents frequent associations between dolphins) Lusseau et al. (2003); and (4) a social network from a high tech firm where no description is available on the network repository. The statistics for the four networks as reported by the Network Repository are provided in Table 1. The θ values used in the viral processes for the networks are: retweet $\theta = .98\lambda_{\max}$, karate $\theta = .96\lambda_{\max}$, dolphins $\theta = .95\lambda_{\max}$, firm-hi-tech $\theta = .99\lambda_{\max}$.

In Fig. 4 we plot $E[N_\infty]$ versus assortativity for the rewired versions of each of our four networks. There is clearly a positive but nonlinear relationship between virality and assortativity. However, for fixed assortativity (and fixed degree distribution by design) the virality models produce $E[N_\infty]$ values differing by an order of magnitude between network rewirings in some cases. This highlights the need for further explanatory variables to explain the virality, such as our proposed graphlet frequency distribution. As an important first check as to whether graphlet frequency distribution might possibly play an important role, we also plot in Fig. 4 the variation in graphlet frequency distribution across the rewired networks, observing that there can be significant differences in graphlet frequency distribution between network rewirings. We emphasize here that the graphlet frequency distribution is a measure that is made on each

Table 1 Network statistics as reported by Rossi and Ahmed (2015)

	rt-retweet	karate	soc-dolphins	soc-firm-hi-tech
Nodes	96	34	62	33
Edges	117	78	159	124.5
Density	0.0257	0.1390	0.0841	0.2358
Maximum degree	17	17	12	28
Minimum degree	1	1	1	0
Average degree	2	4	5	9.19
Assortativity	−0.1792	−0.4756	−0.0436	−0.1200
Number of triangles	36	135	285	454.5
Average number of triangles	0	3	4	13.77
Maximum number of triangles	6	18	17	88.5
Average clustering coefficient	0.0608	0.5706	0.2590	0.4050
Fraction of closed triangles	0.0742	0.2557	0.3088	0.2960
Maximum k-core	4	5	5	8

individual rewiring, and that the plot in Fig. 4 summarizes these various distributions for all of our simulated rewirings. So, for example, g28 in network soc-firm-hi-tech displays a small value within the graphlet frequency distribution, meaning that in any given rewiring, there are relatively few of these graphlets present. However, noting that the spread of values for this graphlet on the plot does not include 0, it can be concluded that every rewiring of this network displayed this graphlet to some extent. Since the box plot displays non-trivial variations of graphlet frequency distribution between network rewirings, it is at least possible that the graphlet frequency distribution could be a contributing factor to the variance of $E[N_\infty]$ observed at fixed assortativity in Fig. 4.

To verify whether graphlet frequency distribution does in fact play an important role in virality, in Table 2 we provide the results for a nested regression predicting the Hawkes virality statistic where assortativity or graphlets alone are used, compared to the full model (Eq. 9) with both assortativity and graphlets. We observe that the R^2 values when using only graphlets are slightly larger than when using only assortativity for all networks but soc-dolphins, where the R^2 for graphlets alone is considerably smaller than that of assortativity alone. We observe that the R^2 values increase by around 10% over assortativity alone when graphlets are also considered, and in all cases but one, including graphlets with assortativity allow for over 90% of the variance to be explained. The mean square error also improves with the addition of graphlets to assortativity in the model, with the improvement being a factor of 2–4. In Tables 3 and 4, we provide the analogous results for a nested regression predicting the SIS statistic and largest eigenvalue (respectively). Here we see similar improvements when the graphlets are added to the regression model over assortativity alone, and find that graphlets alone compare to assortativity alone in a similar manner as in the Hawkes model.

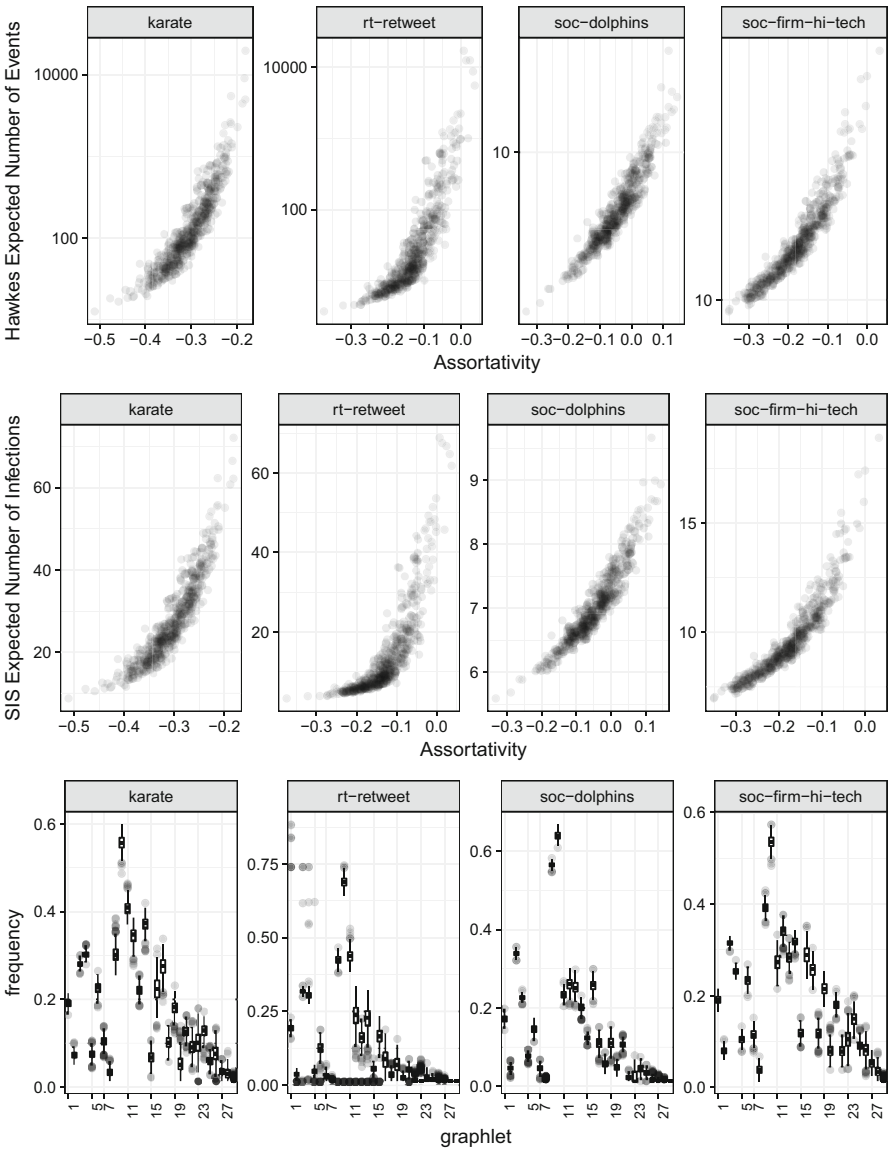

Fig. 4 Top: Hawkes $E[N_\infty]$ versus assortativity in the 500 rewired networks for each real-world network. Middle: SIS $E[N_\infty]$ versus assortativity. Bottom: Box plot of graphlet frequency distributions $d(g_i)$ across the 500 simulated networks for each real-world network

Next we inspect the statistical significance of the regression coefficients to better understand which graphlets are predictive of $E[N_\infty]$ and $\lambda_{\max}$. In Table 5, we list the independent variables in Eq. 9 that are significant at the .01 level. For the rt-retweet and soc-dolphins networks, graphlets lower than g8 are never selected and it appears that larger graphlets are needed to improve the model beyond assortativity. On the other

Table 2 Model comparison for predicting log number of events for Hawkes

Network	Assort		GFD		Assort + GFD	
	R^2	MSE	R^2	MSE	R^2	MSE
rt-retweet	7.22E−01	1.27E−01	7.81E−01	9.96E−02	8.62E−01	6.31E−02
Karate	8.14E−01	3.95E−02	8.97E−01	2.19E−02	9.63E−01	7.94E−03
soc-dolphins	8.65E−01	3.03E−04	7.12E−01	6.43E−04	9.04E−01	2.15E−04
soc-firm-hi-tech	8.54E−01	2.33E−03	8.70E−01	2.08E−03	9.33E−01	1.07E−03

Table 3 Model comparison for predicting log number of events for SIS

Network	Assort		GFD		Assort + GFD	
	R^2	MSE	R^2	MSE	R^2	MSE
rt-retweet	8.36E−01	1.32E−02	8.58E−01	1.14E−02	9.28E−01	5.78E−03
Karate	8.77E−01	2.91E−03	9.08E−01	2.17E−03	9.76E−01	5.59E−04
soc-dolphins	8.86E−01	1.47E−04	7.05E−01	3.80E−04	9.16E−01	1.08E−04
soc-firm-hi-tech	8.93E−01	5.24E−04	8.78E−01	6.00E−04	9.57E−01	2.09E−04

Table 4 Model comparison for predicting log of largest eigenvalue

Network	Assort		GFD		Assort + GFD	
	R^2	MSE	R^2	MSE	R^2	MSE
rt-retweet	8.51E−01	4.21E−05	8.98E−01	2.87E−05	9.50E−01	1.41E−05
Karate	8.66E−01	9.74E−06	9.14E−01	6.28E−06	9.78E−01	1.61E−06
soc-dolphins	8.74E−01	3.98E−06	7.21E−01	8.80E−06	9.15E−01	2.69E−06
soc-firm-hi-tech	9.25E−01	2.82E−06	8.77E−01	4.62E−06	9.72E−01	1.04E−06

hand, lower-order graphlets are significant for the soc-firm-hi-tech network, where triangles are significant across the viral process models.

To improve the regression model in Eq. 9, we consider an interaction model where the statistically significant variables from Table 5 are used and interaction terms of the form $A \cdot \log(d(g_i))$ are added. In Table 5 we display the R^2 values for this interaction model. In some cases, we see large improvements, for example in the case of the retweet network and the Hawkes model the R^2 value increases from .86 to .95 (the R^2 for assortativity alone is .72). In the majority of cases the R^2 value of this interaction model is above .95 and for the karate model is above .98. The R^2 value for the soc-dolphins network is slightly lower, .9–.92. Given that only high-order graphlets are selected in the soc-dolphins network, it may be the case that graphlets beyond g_29 are needed to achieve R^2 values close to 1 for that network.

Table 5 Important variables along with R^2 values for interaction regression model

Network	Important variables (.01 level)	R^2
Hawkes		
rt-retweet	A, g8, g10, g13, g18, g21, g22, g24, g25, g27	0.950
karate	A, g1, g3, g4, g6, g9, g10, g11, g12, g13, g14, g16, g17, g18, g19, g28	0.983
soc-dolphins	A, g10, g12, g25, g26, g28	0.913
soc-firm-hi-tech	A, g1, g3, g4, g6, g9, g10, g11, g12, g13, g14, g16, g17, g18, g19, g23, g27, g28, g29	0.977
SIS		
rt-retweet	A, g10, g11, g14, g15, g16, g17, g19, g20, g21, g23, g24, g25	0.969
karate	A, g3, g10, g12, g14, g15, g18, g19, g20, g24, g26, g27, g28	0.986
soc-dolphins	A, g21, g24, g26	0.902
soc-firm-hi-tech	A, g3, g8, g10, g12, g15, g19, g25, g27, g28	0.972
λ_{max}		
rt-retweet	A, g10, g13, g18, g21, g22, g24, g25, g27	0.940
karate	A, g10, g12, g13, g14, g15, g18, g19, g20, g23, g24, g25, g26, g27, g28	0.984
soc-dolphins	A, g10, g12, g14, g21, g26	0.929
soc-firm-hi-tech	A, g1, g3, g4, g6, g9, g10, g11, g12, g13, g14, g16, g17, g18, g19, g23, g28	0.977

5 Discussion

Our results show that Hawkes and SIS processes on networks with the same degree distribution and assortativity can exhibit very different levels of virality. With the inclusion of graphlets in predictive models of virality, R^2 values improve by 10–20% depending on the network and specific process. These results have implications for prediction of real network viral processes, as current methods are generally based upon degree distribution alone (Zhao et al. 2015; Crane and Sornette 2008). Typically, approximate models of viral processes on networks can be somewhat readily written in terms of degree distribution, which explains the popularity of this basic network measure in understanding these processes. Better approximate models can be formulated by adding assortativity, which may also be plausible from an analytic point of view. Our results show that adding graphlet frequency distribution to these approximate models could be quite valuable. Here it may be beneficial to consider local graphlet statistics (Dave et al. 2017) around the source node of the process. For example, the local graphlet frequency around a Tweeter may help predict the number of re-shares of that user's Tweet. Future work will focus on analytical attempts at capturing the role graphlet frequency distribution plays in these viral processes, now that their role has been experimentally verified.

The statistically significant graphlets for a given network may also provide useful information for network optimization. When the goal is to reduce the spread of viruses, for example to mitigate fake news (Farajtabar et al. 2017), one may wish to remove nodes in a way that leads to the greatest reduction in viral spreading. While degree

may be used to make such selections, further gains may be made by considering assortativity and graphlet frequency. For example, in the case of the dolphins-SIS simulation, choosing nodes that reduce the frequency of graphlets 21, 24, and 26 may provide a better mitigation strategy than other metrics like centrality or degree. These questions will be addressed in future research.

Acknowledgements This work was supported in part by NSF Grants SCC-1737585, SES-1343123, ATD-1737996, and ATD-1737925.

References

Anderson, R.M., May, R.M., Anderson, B.: Infectious Diseases of Humans: Dynamics and Control, vol. 28. Wiley Online Library, New York (1992)

Berger, N., Borgs, C., Chayes, J.T., Saberi, A.: On the spread of viruses on the internet. In: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 301–310. Society for Industrial and Applied Mathematics (2005)

Bhuiyan, M., Rahman, M., Rahman, M., Al Hasan, M.: GUISE: uniform sampling of graphlets for large graph analysis. In: 12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, pp. 91–100 (2012)

Callaway, D.S., Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Network robustness and fragility: percolation on random graphs. Phys. Rev. Lett. **85**(25), 5468 (2000)

Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., Faloutsos, C.: Epidemic thresholds in real networks. ACM Trans. Inf. Syst. Secur. **10**(4), 1:1–1:26 (2008a)

Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., Faloutsos, C.: Epidemic thresholds in real networks. ACM Trans. Inf. Syst. Secur. (TISSEC) **10**(4), 1 (2008b)

Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system. Proc. Natl. Acad. Sci. **105**(41), 15649–15653 (2008)

Dave, V., Ahmed, N., Al Hasan, M.: E-CLoG: counting edge-centric local graphlets. In: Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), BIG DATA '17. IEEE Computer Society (2017)

Dezső, Z., Barabási, A.-L.: Halting viruses in scale-free networks. Phys. Rev. E **65**(5), 055103 (2002)

Dietz, K.: Models for vector-borne parasitic diseases. In: Vito Volterra Symposium on Mathematical Models in Biology, pp. 264–277. Springer, Berlin (1980)

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H.: Fake news mitigation via point process based intervention.In: International Conference on Machine Learning, pp. 1097–1106 (2017)

Ganesh, A., Massoulie, L., Towsley, D.: The effect of network topology on the spread of epidemics. In: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 1455–1466 (2005)

Givan, O., Schwartz, N., Cygelberg, A., Stone, L.: Predicting epidemic thresholds on complex networks: limitations of mean-field approaches. J. Theor. Biol. **288**, 21–28 (2011)

Jalan, S., Yadav, A.: Assortative and disassortative mixing investigated using the spectra of graphs. Phys. Rev. E **91**(1), 012813 (2015)

Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (TWEB) **1**(1), 5 (2007)

Lovasz, L.: Eigenvalues of Graphs. http://web.cs.elte.hu/~lovasz/eigenvals-x.pdf (2007)

Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. **54**(4), 396–405 (2003)

Maslov, S., Sneppen, K.: Specificity and stability in topology of protein networks. Science **296**(5569), 910–913 (2002)

Mihail C.G.M., Zegura, E.: The Markov chain simulation method for generating connected power law random graphs. In: Proceedings of the 5th Workshop on Algorithm Engineering and Experiments (ALENEX). SIAM (2003)

Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)

Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
Qu, J., Wang, S.-J., Jusup, M., Wang, Z.: Effects of random rewiring on the degree correlation of scale-free networks. Sci. Rep. **5**, 15450 (2015)
Rahman, M., Al Hasan, M.: Sampling triples from restricted networks using MCMC strategy. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, pp. 1519–1528 (2014)
Rahman, M., Al Hasan, M.: Link prediction in dynamic networks using graphlet. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 394–409. Springer, Berlin (2016)
Rahman, M., Bhuiyan, M., Al Hasan, M.: GRAFT: an approximate graphlet counting algorithm for large graph analysis. In: 21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, pp. 1467–1471 (2012)
Rahman, M., Bhuiyan, M.A., Al Hasan, M.: GRAFT: an efficient graphlet counting method for large graph analysis. IEEE Trans. Knowl. Data Eng. **26**(10), 2466–2478 (2014a)
Rahman, M., Bhuiyan, M.A., Rahman, M., Al Hasan, M.: GUISE: a uniform sampler for constructing frequency histogram of graphlets. Knowl. Inf. Syst. **38**(3), 511–536 (2014b)
Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
Rossi, R.A., Gleich, D.F., Gebremedhin, A.H., Patwary, M.A.: What if CLIQUE were fast? Maximum Cliques in Information Networks and Strong Components in Temporal Networks, pp. 1–11. arXiv preprint arXiv:1210.5802 (2012)
Saha, T.K., Al Hasan, M.: Finding network motifs using MCMC sampling. In: Complex Networks VI—Proceedings of the 6th Workshop on Complex Networks CompleNet 2015, New York City, pp. 13–24 (2015)
Short, M.B., Mohler, G.O., Brantingham, P.J., Tita, G.E.: Gang rivalry dynamics via coupled point process networks. Discrete Contin. Dyn. Syst. Ser. B **19**(5), 1459–1477 (2014)
Van Mieghem, P., Wang, H., Ge, X., Tang, S., Kuipers, F.A.: Influence of assortativity and degree-preserving rewiring on the spectra of networks. Eur. Phys. J. B Condens. Matter Complex Syst. **76**(4), 643–652 (2010)
Yang, L.-X., Draief, M., Yang, X.: The impact of the network topology on the viral prevalence: a node-based approach. PLoS ONE **10**(7), e0134507 (2015)
Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: Seismic: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1513–1522. ACM (2015)