

# ULRR

## A generating-function approach to modelling complex contagion on clustered networks with multi-type branching processes

Item Type	Article
Authors	Keating, Leah A.; Gleeson, James; O'Sullivan, David
Citation	Journal of Complex Networks 11(6)
Publisher	Oxford University Press
Download date	2026-04-16 07:52:27
Item License	<a href="https://creativecommons.org/licenses/by-nc-sa/4.0/">https://creativecommons.org/licenses/by-nc-sa/4.0/</a>
Link to Item	<a href="https://doi.org/10.34961/researchrepository-ul.24745572">https://doi.org/10.34961/researchrepository-ul.24745572</a>

## A generating-function approach to modelling complex contagion on clustered networks with multi-type branching processes

LEAH A. KEATING<sup>1,2,†</sup> , JAMES P. GLEESON<sup>1</sup> , AND DAVID J. P. O'SULLIVAN<sup>1</sup> 

<sup>1</sup>MACSI, Department of Mathematics and Statistics, University of Limerick,  
Limerick V94 T9PX, Ireland

<sup>2</sup>Department of Mathematics, University of California, Los Angeles, CA 90095, USA

†Corresponding author. Email: leahkeating@math.ucla.edu

[Received on 26 June 2023; editorial decision on 2 October 2023; accepted on 12 October 2023]

Understanding cascading processes on complex network topologies is paramount for modelling how diseases, information, fake news and other media spread. In this article, we extend the multi-type branching process method developed in Keating *et al.*, (2022), which relies on networks having homogenous node properties, to a more general class of clustered networks. Using a model of socially inspired complex contagion we obtain results, not just for the average behaviour of the cascades but for full distributions of the cascade properties. We introduce a new method for the inversion of probability generating functions to recover their underlying probability distributions; this derivation naturally extends to higher dimensions. This inversion technique is used along with the multi-type branching process to obtain univariate and bivariate distributions of cascade properties. Finally, using clique-cover methods, we apply the methodology to synthetic and real-world networks and compare the theoretical distribution of cascade sizes with the results of extensive numerical simulations.

**Keywords:** network dynamics, branching processes, complex contagion, probability-generating functions

### 1. Introduction

Recent developments in technology have led to social network data becoming more available and more analysed than ever before. For example, the Twitter API V2 allows academic researchers to access up to 10 million tweets per month. With this wealth of social network data, we need to develop appropriate tools for its analysis. In this article, we focus on learning about the interplay between the network structure and the dynamics on the network, such as the spread of behaviour, information or a disease. Social networks are highly clustered [1–4], this means that they contain a higher number of triangles than random networks, reflecting the fact that *a friend of my friend is likely to be a friend of mine*; however, most models used to model diffusion in online social networks assume that the network is locally tree-like and thus ignoring the clustering [2, 5–7]. It is important that we develop tools for analysing network dynamics which account for this clustering.

In an online experiment, Centola [8] observed that repeated exposures to a health behaviour from network neighbours made its adoption more likely, given that the individual had not already adopted. This adoption mechanism is known as a *complex contagion*. Complex-contagion dynamics have also been observed in the use of Skype add-ons [9], the spread of information and politically controversial hashtags on Twitter [10, 11] and the spread of online fads [12]. Clustering in the underlying network has been shown to inhibit simple-contagion dynamics [13], this is because the clustering leads to redundancy in the links for transmission; however, clustering drives the reinforcement mechanism of the complex contagion [8, 14] and therefore larger outbreaks can occur in clustered networks for a complex contagion.

Most commonly, complex contagions are modelled using threshold models [15]. In a such model, each node is assigned a threshold and when the number of neighbours who have adopted exceeds the threshold, the node adopts with certainty. In Keating *et al.* [16], we introduced an alternative model for complex contagion which is an extension of the independent cascade model (ICM) [17]. We continue to use this model here as it can be easily incorporated into the branching process framework. It allows us to control the effects of social reinforcement through a single parameter,  $\alpha \in [0, 1]$ , and if we set  $\alpha = 0$  we get the ICM of [17] which is a simple contagion; that is we can study both simple and complex contagions through this model.

Branching processes are discrete-time stochastic processes that have been used to model information cascades [18–23], the spread of infectious diseases [24, 25] and population dynamics in ecology [26]. Simple branching processes have proven a fruitful means of capturing important network properties. Gleeson *et al.* [6] analytically derived cascade properties from data including the cascade size distribution, cascade lifetime distribution, expected average tree depth (EATD) and structural virality [27]. The authors focused on the case where the network is assumed to be locally tree-like; that is, unclustered, and considered simple-contagion dynamics only. This also means that we assume that the network has no cycles of length greater than 3. The branching process theory for modelling cascades on networks assumes an infinitely large network; however, as has been shown in numerous previous articles [6, 16, 28] and in this article, branching processes can give very accurate results even when the network size is clearly finite. In previous work [16], we developed a method for modelling complex contagion on clustered networks using multi-type branching processes (MTBPs). We concentrated on average measures of the cascades and analytically calculated the cascade condition and the expected cascade size for a very specific class of networks. In this article, for clustered networks, we analytically derive results including the complete distribution of cascade sizes, the distribution of cascade lifetimes and the EATD. We also extend the MTBP theory to more general network distributions, making the method more applicable to real-world networks. In particular, we use a family of network distributions proposed independently by Newman [29] and Miller [30] in 2009; described by the distribution of triangle and single-link membership of the nodes, we refer to these networks as Newman–Miller networks. By incorporating clustering into the model through the Newman–Miller networks, we can analytically study both simple- and complex-contagion spreading processes.

Within the branching process framework, we use probability generating functions (PGFs) to derive distributions of the quantities of interest. When we have a PGF for a quantity such as cascade size, we are interested in recovering its full probability distribution from the PGF. The most common method of numerically finding the probability distribution in the network dynamics literature is that of Cavers [22, 28, 31], this involves using the z-transform inversion which has the form of a contour integral and approximating the integral using the trapezoidal rule. For bivariate PGFs, Brummitt *et al.* [32] use three different methods of finding the probability distribution in a two-type branching process, but these methods use computer algebra systems and are restricted in the number of terms that they can retrieve. While Cavers' derivation has been extended to bivariate PGFs [33], it is not clear how it extends to higher dimensions. In this article, we introduce an alternative derivation of the PGF inversion method. Our approach differs to that of Cavers [31] and Antal [33]; while they focus on the z-transform inversion, we look at the equivalence between the PGF evaluated at specific points and the discrete Fourier transform of the probability distribution. Our method has the same computational implementation as Cavers' method in one dimension but has the added advantage of naturally extending to the inversion of PGFs of any dimension. One of the contributions of this article is to describe and show how to implement the inversion of univariate and bivariate PGFs, the same method may be straightforwardly extended to higher-dimensional distributions.

We are interested in understanding how the structure of a network and the dynamics interact with each other. It is important that we can apply these methods to real-world networks. Burgio *et al.* [34] introduced a clique-cover method which allows us to approximate the (maximal) clique-membership distribution of a network while constraining that the cliques are edge-disjoint—cliques do not share edges—called the edge-disjoint edge clique cover (EECC). In Section 6 we apply the EECC to synthetic and real-world networks and find the cascade size distribution according to the branching process theory, for comparison with Monte-Carlo simulations. Applying the MTBP to these empirical networks allows us to see where the theory gives accurate results and to identify areas for future work on the branching process methodology. When we approximate a network using the EECC, we lose some information about the network structure. Some larger cliques may be approximated as a group of independent smaller cliques and we also lose information on correlations in the network structure. However, despite this information loss, the method is accurate for many of the examples in this article.

The rest of this article is structured as follows; in Section 2, we describe the complex contagion adoption dynamics, in Section 3, we introduce the class of networks that we use in our calculations, in Section 4, we show how the MTBP framework can be leveraged to derive distributions of cascade properties. Our examples in Section 4 include the distributions of cascade size and cascade lifetimes and the joint distribution of cascade size and cumulative depth. In Section 5, we derive a method for the inversion of PGFs to recover the probability distribution, in Section 6, we discuss and show results for applying this method to real-world, networks, and in Section 7, we discuss the capabilities and limitations of our approach.

## 2. Complex contagion adoption dynamics

When we study a spreading process on a network, both the spreading mechanism and the network structure are required to accurately model the process. To model complex contagion using the MTBP methodology, we use the discrete-time model of [16], which is an extension of the ICM [17]. In this model, nodes in the network are assumed to be in one of three states; active, inactive or removed. The spreading process is initiated with a small number of active seed nodes that are randomly selected. In examples here we choose to select just one seed node, all other nodes are initially inactive. At each time step, active nodes have one chance to activate each of their inactive neighbours, activation occurs with probability  $p_k$ , where  $k$  is the number of times that that inactive neighbour has been exposed. If the activation is successful, the inactive neighbour becomes active in the next time step, active nodes always become removed in the following time step and once a node enters the removed state, it never leaves that state. The dynamics are governed by two parameters;  $p_1$  and  $\alpha$ , where  $p_1 \in [0, 1]$  is the probability of adoption after a single exposure and  $\alpha \in [0, 1]$  is a measure of the strength of social reinforcement. Higher values of  $\alpha$  represent higher levels of social reinforcement and in the limiting case of  $\alpha = 0$ , we recover exactly the (simple-contagion) independent cascade model. In the limiting case of  $\alpha = 1$ , adoption occurs with certainty on the  $k$ th exposure when  $k \geq 2$  or with probability  $p_1$  on the first exposure; that is, if  $k = 1$ . The probability that a node will become active after its  $k$ th exposure, given that it is inactive, is given by the following equations,

$$p_k = 1 - q_k \tag{2.1}$$

and

$$q_k = q_1(1 - \alpha)^{k-1}, \tag{2.2}$$

where  $q_k$  is the probability that a node does not adopt directly after the  $k$ th exposure, given that it has not already adopted. We studied this model in [16] for a small number of networks with homogeneous Newman–Miller degree distributions. Some of the results included the cascade condition and the expected cascade size.

### 3. Newman–Miller distributions

Along with the complex contagion adoption dynamics that we described in the previous section, to fully describe the MTBP we need a way of describing the network so that it can be incorporated into the MTBP. We model diffusion on networks where the clique membership of the nodes follows a joint probability distribution  $\pi_{st}$ , which is defined as the probability that a node, chosen at random, is in  $s$  single links and  $t$  triangles (is in  $s$  2-cliques and  $t$  3-cliques). We refer to  $s$  as the *link degree* and  $t$  as the *triangle degree*. This family of networks was independently proposed by Newman [29] and Miller [30] in 2009, and we have elected to use these network models because they are commonly used to analytically study spreading processes on clustered networks [23, 35]. We use PGFs to describe the joint triangle-link (or Newman–Miller) degree distribution of the network. The PGF is defined as

$$\tilde{f}(x, y) = \sum_{s,t} \pi_{st} x^s y^t. \quad (3.1)$$

Here, we use the example of the doubly Poisson distribution [29] which has probability mass function (PMF)

$$\pi_{st} = e^{-\mu} \frac{\mu^s}{s!} e^{-\nu} \frac{\nu^t}{t!}, \quad (3.2)$$

where  $\mu$  is the expected number of single links per node and  $\nu$  is the expected number of triangles per node. We can express this distribution in the form of a PGF using eq. (3.1),

$$\tilde{f}(x, y) = \sum_{s,t=0}^{\infty} e^{-\mu} \frac{\mu^s}{s!} e^{-\nu} \frac{\nu^t}{t!} x^s y^t = e^{\mu(x-1) + \nu(y-1)}. \quad (3.3)$$

When we explore the dynamics of the system, we often need to know the excess degree distributions from traversing a link or a triangle, to allow us to calculate the cascade size distribution, for example. If we traverse a single link, then the joint excess degree is the number of triangles and single links that the node at the other end is a part of, not including the single link traversed itself. If we traverse a triangle, the joint excess degree is the number of single links and triangles that the node at the other end is part of, not including the triangle traversed. As described by Newman [29], we can find the joint excess triangle-link distributions from traversing a single link  $f_q$  and from traversing a triangle  $f_r$  as follows;

$$f_q = \frac{\sum_{st} s \pi_{st} x^{s-1} y^t}{\sum_{st} s \pi_{st}} = \frac{\partial / \partial x \tilde{f}(x, y)}{\partial / \partial x \tilde{f}(x, y)|_{x=y=1}} \quad (3.4)$$

and

$$f_r = \frac{\sum_{st} t \pi_{st} x^s y^{t-1}}{\sum_{st} t \pi_{st}} = \frac{\partial / \partial y \tilde{f}(x, y)}{\partial / \partial y \tilde{f}(x, y)|_{x=y=1}}. \quad (3.5)$$

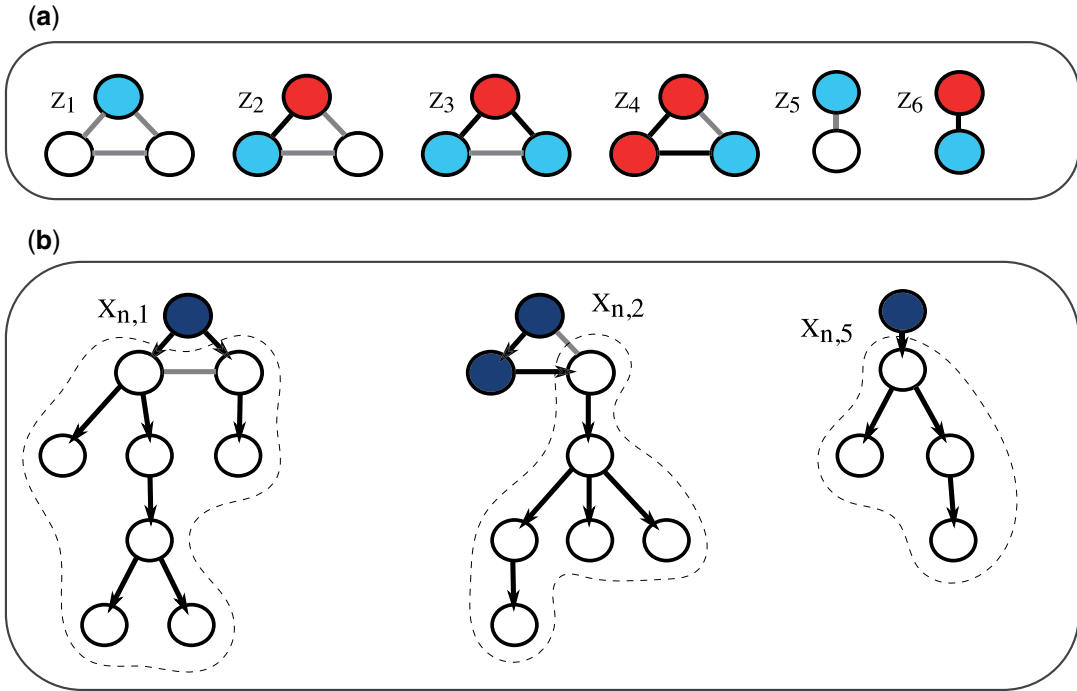


FIG. 1. (a) The six possible clique motifs in a Newman–Miller network, blue nodes are active, white nodes are inactive, and red nodes are removed. (b) A visualization of the node counting in subtrees for cascade size, we do not count active or removed nodes in the seed motif (dark blue) and only count all other nodes in the subtree (in white). In this figure, we assume that the number of generations that we allow the subtree to grow for  $n \rightarrow \infty$ . We let  $n \rightarrow \infty$  because, for a sub-critical branching process, this assures that the sub-trees will be fully grown; that is, there will be no further offspring in future generations.

In the following sections, we use these network-structure distributions in our calculations of the distributions and values of quantities of interest.

#### 4. Derivations of cascade properties

In this section, using PGFs, we theoretically derive distributions of cascade size, cascade lifetime and the joint distribution of cascade size and cumulative depth. We also use the moments of the joint distribution to calculate the Pearson’s correlation coefficient between cascade size and cumulative depth and the expected average tree depth (EATD).

##### 4.1 Cascade size distribution

Here, we analytically derive the cascade-size distribution for complex-contagion dynamics on a network constructed using the Newman–Miller approach described in Section 3. Previously, we have calculated the average cascade behaviour using MTBPs [16]. The *types* in the MTBP are *clique motifs*, as shown in Fig. 1(a), a clique motif is a clique with a specific number of active, inactive and removed nodes. We are only concerned with those motifs with at least one active node. In a Newman–Miller network there are six possible clique motifs, we denote by  $z_i$  a type- $i$  motif. In the article [16], we showed how

to calculate the expected cascade size; however, often it is the distribution of cascade sizes that is of interest, not just the expected value. Other works analytically derive this distribution for locally tree-like networks [6, 21, 22] but not for clustered networks. Here, we derive the distribution of cascade sizes for the complex-contagion dynamics that we described in Section 2 on (clustered) Newman–Miller networks.

We consider the size of a full cascade after  $n$  generations,  $\tilde{X}_n$ , in terms of its subtree sizes, where the size of a type- $i$  subtree after  $n$  generations,  $X_{n,i}$ , is the number of nodes in the tree seeded by a type- $i$  motif that are currently active or removed, excluding those active or removed nodes that were in the seed motif; this is illustrated in Fig. 1(b). Not including these active or removed nodes in the seed motif avoids counting a node in the full cascade more than once.

To find the cascade-size distribution, we start by finding the PGF for cascade size. By definition, the PGF for cascade size for a full tree that has grown from its seed for  $n$  generations,  $\tilde{K}_n(z)$ , is

$$\tilde{K}_n(z) = \sum_{l=0}^{\infty} P(\tilde{X}_n = l) z^l = \mathbb{E} \left[ z^{\tilde{X}_n} \right], \quad (4.1)$$

where  $P(\tilde{X}_n = l)$  is the probability that the size of the full cascade  $\tilde{X}_n$  is  $l$ . The full cascade size,  $\tilde{X}_n$ , can be expressed in terms of its subtree sizes,

$$\tilde{X}_n = 1 + \sum_{i=1}^s X_{n,5}^{(i)} + \sum_{j=1}^t X_{n,1}^{(j)}, \quad (4.2)$$

where  $s$  is the number of subtrees seeded with a type-5 motif and  $t$  is the number of subtrees seeded with a type-1 motif—the seed node is in  $t$  triangles and has  $s$  single links (see Fig. 1(a) for the motif types). The PGF for the full cascade size after  $n$  generations,  $\tilde{K}_n(z)$ , is given by

$$\begin{aligned} \tilde{K}_n(z) &= \mathbb{E} \left[ z^{\tilde{X}_n} \right] \\ &= \sum_{s,t} \pi_{st} \mathbb{E} \left[ z^{1 + \sum_{i=1}^s X_{n,5}^{(i)} + \sum_{j=1}^t X_{n,1}^{(j)}} \middle| \text{seed node has } s \text{ links and } t \text{ triangles} \right] \\ &= z \sum_{s,t} \pi_{st} \mathbb{E} \left[ z^{X_{n,5}} \right]^s \mathbb{E} \left[ z^{X_{n,1}} \right]^t \\ &= z \tilde{f} \left( K_{n,5}(z), K_{n,1}(z) \right). \end{aligned} \quad (4.3)$$

To find  $\tilde{K}_n(z)$ , we need to know the PGFs for the subtree sizes  $K_{n,1}(z)$ ,  $K_{n,2}(z)$  and  $K_{n,5}(z)$ . The PGF for a type-1 subtree, after  $n$  generations is given by,

$$\begin{aligned} K_{n,1}(z) &= \mathbb{E} \left[ z^{X_{n,1}} \right] \\ &= \sum_{l=0}^2 \binom{2}{l} p_1^l (1-p_1)^{2-l} \mathbb{E} \left[ z^{X_{n,1}} \mid l \text{ active nodes in generation 1} \right] \\ &= (1-p_1)^2 z^0 \end{aligned} \quad (4.4)$$

$$\begin{aligned}
 &+ 2p_1(1 - p_1)\mathbb{E} \left[ z^{1+X_{n-1,2}+\sum_i X_{n-1,1}^{(i)}+\sum_j X_{n-1,5}^{(j)}} \right] \\
 &+ p_1^2\mathbb{E} \left[ z^{2+\sum_i X_{n-1,1}^{(i)}+\sum_j X_{n-1,5}^{(j)}+\sum_k X_{n-1,1}^{(k)}+\sum_l X_{n-1,5}^{(l)}} \right],
 \end{aligned}$$

where  $X_{n,1}$  is a subtree seeded by a type-1 motif. The three terms in eq. (4.4) come from the fact that a type-1 motif, in the next generation, either gives rise to no new active nodes, one new active node or two new active nodes, respectively. These new active nodes must be counted in the cascade size along with the generation  $n - 1$  subtrees that they seed. For example, if in generation 1 a type-1 motif has 1 offspring; that is, produces a type-2 motif, the subtree size can be written in the form,  $1 + X_{n-1,2} + \sum_i X_{n-1,1}^{(i)} + \sum_j X_{n-1,5}^{(j)}$ , where the 1 counts the new active node,  $X_{n-1,2}$  is for the size of the type-2 subtree produced and the  $\sum_i X_{n-1,1}^{(i)} + \sum_j X_{n-1,5}^{(j)}$  is for the generation  $n - 1$  type-1 and type-5 subtrees produced, where the number of type-1 subtrees is the excess triangle degree from traversing a triangle and the number of type-5 subtrees is the excess link degree from traversing a triangle. We can rewrite eq. (4.4) as

$$\begin{aligned}
 K_{n,1}(z) &= (1 - p_1)^2 \tag{4.5} \\
 &+ 2p_1(1 - p_1)z\mathbb{E} \left[ z^{X_{n-1,2}} \sum_{s,t} r_{st}\mathbb{E} \left[ z^{X_{n-1,5}} \right]^s \left[ z^{X_{n-1,1}} \right]^t \right] \\
 &+ p_1^2z^2 \left( \sum_{s,t} r_{st}\mathbb{E} \left[ z^{X_{n-1,5}} \right]^s \left[ z^{X_{n-1,1}} \right]^t \right)^2,
 \end{aligned}$$

which becomes

$$\begin{aligned}
 K_{n,1}(z) &= (1 - p_1)^2 + 2p_1(1 - p_1)zK_{n-1,2}(z)f_r(K_{n-1,5}(z), K_{n-1,1}(z)) \\
 &+ p_1^2z^2 \left( f_r(K_{n-1,5}(z), K_{n-1,1}(z)) \right)^2
 \end{aligned}$$

when we notice that these expectations and sums are equivalent to the PGFs for the offspring distribution from traversing a triangle,  $f_r(x, y)$ , and the PGFs for subtree size after  $n - 1$  generations. Similarly, by using the offspring distributions from a type-2 and a type-5 motif we can get the subtree size PGFs for type-2 and type-5 subtrees:

$$\begin{aligned}
 K_{n,2}(z) &= \mathbb{E} \left[ z^{X_{n,2}} \right] \tag{4.6} \\
 &= 1 - p_2 + p_2zf_r(K_{n-1,5}(z), K_{n-1,1}(z)),
 \end{aligned}$$

and

$$\begin{aligned}
 K_{n,5}(z) &= \mathbb{E} \left[ z^{X_{n,5}} \right] \tag{4.7} \\
 &= 1 - p_1 + p_1zf_q(K_{n-1,5}(z), K_{n-1,1}(z)).
 \end{aligned}$$

Initially (in generation  $n = 0$ ) the subtree size is zero; that is,  $X_{0,1} = X_{0,2} = X_{0,5} = 0$ , giving us initial conditions for their respective PGFs of  $\mathbb{E} \left[ z^{X_{0,i}} \right] = \mathbb{E} \left[ z^0 \right] = 1$ . We iterate through eqns. (4.3) to (4.7) with initial conditions;  $K_{0,1}(z) = K_{0,2}(z) = K_{0,5}(z) = 1$ , to find  $\tilde{K}_n(z)$ . With this, we have a system of equations

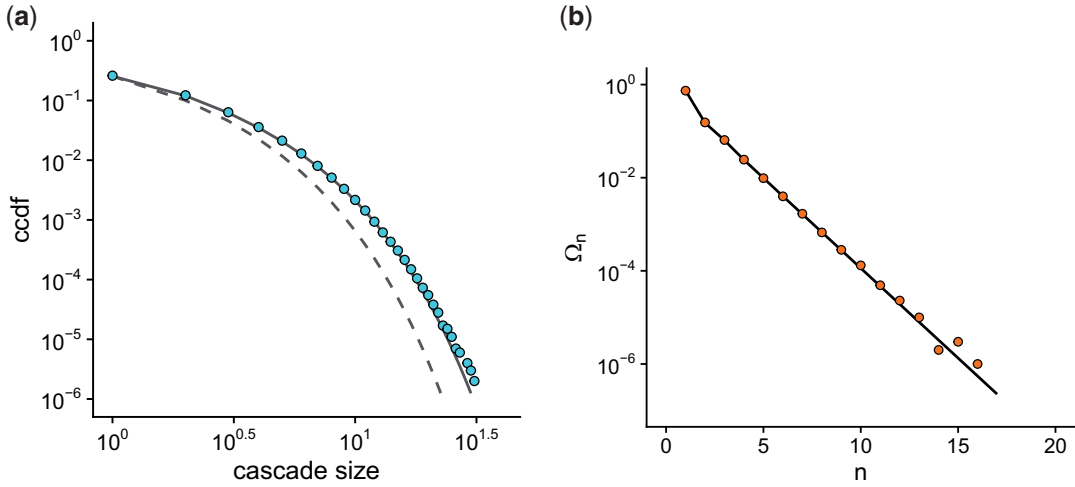


FIG. 2. (a) The theoretical complementary cumulative distribution function (CCDF) of cascade size from the MTBP (solid line) and for a simple branching process (dashed line) and 1 million simulations on a network of size 10,000 (points). The CCDF for a random variable  $Y$  is defined as  $P(Y > y)$ , the probability that the random variable  $Y$  is greater than a specific  $y$ . (b) The probability distribution of cascade lifetimes where  $\Omega_n$  is the probability that the cascade survives to generation  $n$ . The solid line represents the theoretical distribution and the points are the values from simulations. The results shown are for doubly Poisson Newman–Miller networks with  $\mu = 2$  and  $\nu = 2$  (average degree 6) and for parameters  $p_1 = 0.05$  and  $\alpha = 0.2$ .

that describe the PGF for cascade size; in Section 5, we show how to recover the probability distribution from the PGF, allowing us to plot the cascade-size distribution for a given Newman–Miller network with given values for  $p_1$  and  $\alpha$  in the adoption dynamics. In Fig. 2, we show an example of the cascade size distribution derived from the PGF compared to simulations on a network with a doubly-Poisson distribution. In [16], we calculated the expected cascade size for sub-critical cascades, here we calculate the distribution of cascade sizes for sub-critical cascades in the limit as the number of generations  $n \rightarrow \infty$ . The cascades are *sub-critical* if an infinitesimally small seed fraction of active nodes will not generate cascades of non-finite mean size, which is a non-vanishing fraction of the network as the network size tends to the large network limit. To find the PGF in the limit as  $n \rightarrow \infty$ , we set a tolerance of  $10^{-5}$  such that we stop iterating once the PGF evaluated at generation  $n$  is within  $10^{-5}$  of the PGF evaluated at generation  $n - 1$ ; that is,  $|G_n(x) - G_{n-1}(x)| < 10^{-5}$  for a known number  $x$ .

#### 4.2 Distribution of cascade lifetimes

Here, we show how to find the distribution of cascade lifetimes. The lifetime of a cascade is the number of generations it lives for before dying out; that is, the lifetime is the maximum of the generations of all nodes present in the cascade. If a cascade has particles—by particle we mean an active node—in generation  $n - 1$  but no particles in generation  $n$ , the lifetime of the cascade is  $n$ . To calculate the probability  $\Omega_n$  that the lifetime of a cascade is exactly  $n$  generations, we therefore find the probability that the number of particles in generation  $n$  is zero and subtract the probability that there are zero particles in generation  $n - 1$ . Let the number of particles in generation  $n$  be given by  $\tilde{Z}_n$ , and define the corresponding PGF to be

$$\tilde{F}_n(x) = \sum_{l=0}^{\infty} P(\tilde{Z}_n = l)x^l = \mathbb{E} \left[ x^{\tilde{Z}_n} \right]. \quad (4.8)$$

We write the PGF in terms the number of particles in generation  $n$  of its subtrees

$$\begin{aligned}\tilde{F}_n(x) &= \sum_{s,t} \pi_{st} \mathbb{E} \left[ x^{\sum_{i=1}^s Z_{n,5}^{(i)} + \sum_{j=1}^t Z_{n,1}^{(j)}} \mid \text{seed node has } s \text{ links and } t \text{ triangles} \right] \\ &= \sum_{s,t} \pi_{st} \mathbb{E} [x^{Z_{n,5}}]^s \mathbb{E} [x^{Z_{n,1}}]^t \\ &= \tilde{f}(F_{n,5}(x), F_{n,1}(x)),\end{aligned}\tag{4.9}$$

for  $n \geq 1$ , where  $F_{n,5}(x)$  and  $F_{n,1}(x)$  are the PGFs for the number of particles in generation  $n$  for type-5 and type-1 subtrees, respectively. For completeness, if  $n = 0$  then  $\tilde{F}_0(x) = \mathbb{E} [x^{\tilde{Z}_0}] = \mathbb{E} [x^1] = x$ . To fully compute  $F_n(x)$ , we need  $F_{n,1}(x)$ ,  $F_{n,2}(x)$  and  $F_{n,5}(x)$ , which we calculate in a similar way to the PGFs for subtree size in Section 4.1,

$$\begin{aligned}F_{n,1}(x) &= (1 - p_1)^2 + 2p_1(1 - p_1)F_{n-1,2}(x)f_r(F_{n-1,5}(x), F_{n-1,1}(x)) \\ &\quad p_r^2 f_r^2(F_{n-1,5}(x), F_{n-1,1}(x)),\end{aligned}\tag{4.10}$$

$$F_{n,2}(x) = 1 - p_2 + p_2 f_r(F_{n-1,5}(x), F_{n-1,1}(x))\tag{4.11}$$

and

$$F_{n,5}(x) = 1 - p_1 + p_1 f_q(F_{n-1,5}(x), F_{n-1,1}(x)),\tag{4.12}$$

with the initial conditions:

$$F_{1,1}(x) = (1 - p_1)^2 + 2p_1(1 - p_1)x + p_1^2 x^2,\tag{4.13}$$

$$F_{1,2}(x) = 1 - p_2 + p_2 x\tag{4.14}$$

and

$$F_{1,5}(x) = 1 - p_1 + p_1 x.\tag{4.15}$$

The initial conditions in eqs. (4.13) to (4.15) are derived from the probability distribution of offspring in the next generation for type-1, type-2 and type-5 motifs respectively. For example, we get eq. (4.13) because there are no offspring with probability  $(1 - p_1)^2$ , one offspring with probability  $2p_1(1 - p_1)$  and two offspring with probability  $p_1^2$ . In the initial condition, each of these probabilities is multiplied by  $x^0$ ,  $x^1$  and  $x^2$ , respectively, giving us  $\mathbb{E} [X^{Z_{1,1}}]$  which is the PGF for the number of particles in generation 1 of a type-1 subtree  $F_{1,1}(x)$ . To find  $\Omega_k$ , the probability that a cascade has lifetime  $k$ , we use the property of PGFs that the probability that there are zero particles in generation  $n$  is  $\tilde{F}_n(0)$  and thus

$$\Omega_k = \tilde{F}_k(0) - \tilde{F}_{k-1}(0).\tag{4.16}$$

In Fig. 2(b), we show the distribution of cascade lifetimes for a doubly Poisson distributed network comparing this theory to simulations.

### 4.3 Joint distributions

Many interesting and practically relevant properties of cascades depend on joint distributions of several quantities, and so it is important to move beyond univariate PGFs (as in Section 4.1) to develop methods for inverting multivariate PGFs. For example, we may wish to calculate the expected average tree depth (EATD), which we can find from the joint distribution of cascade size and cumulative depth. The *cumulative depth* of a cascade is the sum of the depths of all of the nodes in the cascade. The average tree depth is the average depth of a node in the tree and the EATD is the expectation of this quantity over the ensemble of cascades. In a cascade, the average tree depth is given by the relationship

$$\text{average tree depth} = \frac{\sum_{i \in V} \text{depth of node } i}{|V|} = \frac{\text{cumulative depth}}{\text{cascade size}}, \quad (4.17)$$

where  $V$  is the set of all nodes in the cascade and  $|V|$  is the size of the set  $V$ ; that is, the number of nodes in the cascade. To calculate the EATD, we need the joint distribution of cumulative depth and cascade size. We denote by  $\tilde{X}_n$  the size of a cascade after  $n$  generations and by  $\tilde{Y}_n$  the cumulative depth of a cascade after  $n$  generations. The joint PGF of cascade size and cumulative depth is

$$\tilde{H}_n(x, y) = \mathbb{E} \left[ x^{\tilde{X}_n} y^{\tilde{Y}_n} \right]. \quad (4.18)$$

We can write cascade size  $\tilde{X}_n$  and cumulative depth  $\tilde{Y}_n$  after  $n$  generations in terms of the size and cumulative depth of the types 1 and 5 subtrees. A type- $i$  subtree is a subtree seeded by a type- $i$  motif, all six motif types are shown in Fig. 1(a). The sizes are denoted by  $X_{n,1}$  and  $X_{n,5}$ , respectively and the depths are denoted by  $Y_{n,1}$  and  $Y_{n,5}$ , respectively. In terms of subtree sizes, the cascade size is

$$\tilde{X}_n = 1 + \sum_{i=1}^t X_{n,1}^{(i)} + \sum_{j=1}^s X_{n,5}^{(j)}, \quad (4.19)$$

where the seed node is in  $t$  triangles and has  $s$  single links and the cumulative depth is

$$\tilde{Y}_n = \sum_{i=1}^t Y_{n,1}^{(i)} + \sum_{j=1}^s Y_{n,5}^{(j)}. \quad (4.20)$$

This allows us to rewrite the PGF in terms of the joint PGFs of its type-1 and type-5 subtrees;

$$\begin{aligned} \tilde{H}_n(x, y) &= \mathbb{E} \left[ x^{1 + \sum_i X_{n,1}^{(i)} + \sum_j X_{n,5}^{(j)}} y^{\sum_i Y_{n,1}^{(i)} + \sum_j Y_{n,5}^{(j)}} \right] \\ &= \sum_{s,t} \pi_{st} \mathbb{E} \left[ x^{1 + \sum_{i=1}^t X_{n,1}^{(i)} + \sum_{j=1}^s X_{n,5}^{(j)}} y^{\sum_{i=1}^t Y_{n,1}^{(i)} + \sum_{j=1}^s Y_{n,5}^{(j)}} \middle| \text{seed in } t \text{ triangles, } s \text{ links} \right] \\ &= x \sum_{s,t} \pi_{st} \mathbb{E} \left[ x^{X_{n,5}} y^{Y_{n,5}} \right]^s \mathbb{E} \left[ x^{X_{n,1}} y^{Y_{n,1}} \right]^t \\ &= x \tilde{f} \left( H_{n,5}(x, y), H_{n,1}(x, y) \right), \end{aligned} \quad (4.21)$$

where  $\tilde{f}(x, y)$  is the PGF for the joint Newman–Miller distribution of the network, and  $H_{n,k}(x, y)$  is the PGF for the joint distribution of cascade size and cumulative depth for a subtree of type  $k$  allowed to grow for  $n$  generations. When calculating the cascade size of a type- $k$  subtree,  $X_{n,k}$ , similarly to in Section 4.1, we do not count the active or removed nodes in the seed motif because this would lead to counting some nodes multiple times when we consider ensembles of subtrees in the overall size. To calculate the full PGF, we will need expressions for  $H_{n,1}(x, y)$ ,  $H_{n,2}(x, y)$  and  $H_{n,5}(x, y)$ , using a method similar to that for the subtree size PGFs in Section 4.1 we get,

$$\begin{aligned} H_{n,1}(x, y) &= \mathbb{E} [x^{X_{n,1}} y^{Y_{n,1}}] \\ &= (1 - p_1)^2 + 2p_1(1 - p_1)xyH_{n-1,2}(xy, y)f_r(H_{n-1,5}(xy, y), H_{n-1,1}(xy, y)) \\ &\quad + p_1^2x^2y^2(f_r(H_{n-1,5}(xy, y), H_{n-1,1}(xy, y)))^2, \end{aligned} \quad (4.22)$$

$$\begin{aligned} H_{n,2}(x, y) &= \mathbb{E} [x^{X_{n,2}} y^{Y_{n,2}}] \\ &= 1 - p_2 + p_2xyf_r(H_{n-1,5}(xy, y), H_{n-1,1}(xy, y)) \end{aligned} \quad (4.23)$$

and

$$\begin{aligned} H_{n,5}(x, y) &= \mathbb{E} [x^{X_{n,5}} y^{Y_{n,5}}] \\ &= 1 - p_1 + p_1xyf_q(H_{n-1,5}(xy, y), H_{n-1,1}(xy, y)), \end{aligned} \quad (4.24)$$

where  $f_r$  and  $f_q$  are the PGFs for the excess Newman–Miller distribution from a triangle and a single link respectively, these are given in eqs. (3.4) and (3.5). We can find the full joint PGF of cascade size and cumulative depth by iterating through these equations from  $n = 0$  with initial conditions  $H_{0,1}(x, y) = H_{0,2}(x, y) = H_{0,5}(x, y) = 1$ . To find the joint probability distribution from the joint PGF, we use an inverse fast Fourier transform method which we describe next in Section 5. While methods that use computer algebra systems have been used in [32], we are not aware of other work on cascade dynamics that uses this method to recover the joint probability distribution of cascade properties. The main issue with using computer algebra systems to calculate the probabilities from PGFs is that they do not allow us to symbolically calculate the PGF for a large number of generations. This limits us in the number of sizes that we can calculate the probabilities for and also reducing the accuracy in the approximation in the limit as the number of generations  $n \rightarrow \infty$ . This has motivated us to propose the inverse fast Fourier method of recovering the probability distribution which we describe next in Section 5, which does not have the same shortcomings as the aforementioned methods.

## 5. Inverting multivariate PGFs to recover probabilities

Given a PGF (or at least an iterative method for calculating it), in many cases, we would like to access the probability distribution of the quantity that it describes. For one-dimensional PGFs, in the network dynamics literature, the most common inversion method used is that of Cavers [31]. In Cavers' method, a link between the PGF and the  $z$  transform is utilized and the inversion is expressed in the form of an inverse  $z$  transform. The inverse  $z$  transform requires the computation of a contour integral in the complex plane, which is approximated using the trapezoidal rule. The trapezoidal-rule approximation is equivalent to taking the inverse discrete Fourier transform of the PGF evaluated at  $K$  evenly spaced points

around the unit circle centred at zero on the complex plane. In practice, we use the inverse fast Fourier transform, which is a fast, efficient algorithm for computing inverses of discrete Fourier transforms, in the place of the inverse discrete Fourier transform. Here, we introduce an alternative derivation which more naturally extends to the inversion of higher-dimensional PGFs. Our method has the same computational implementation as Cavers' method in one dimension. Given the PGF

$$B(z) = \sum_{k=0}^{\infty} p_k z^k, \quad (5.1)$$

we aim to find the probabilities  $\{p_k\}$ ,  $k \in \{0, 1, 2, \dots\}$ , where  $p_k$  is the probability that the random variable, with probability distribution described by the PGF  $B(z)$ , is equal to  $k$ . The discrete Fourier transform of the first  $K$  probabilities is denoted by the set  $\{P_l\}$ ,  $l \in \{0, 1, \dots, K-1\}$  and is given by the relationship

$$P_l = \sum_{k=0}^{K-1} p_k (e^{-\frac{2\pi i}{K} l})^k \approx B(e^{-\frac{2\pi i}{K} l}), \quad (5.2)$$

which is approximately the PGF  $B(z)$  evaluated at  $e^{-\frac{2\pi i}{K} l}$  for  $K$  sufficiently large and  $l \in \{0, 1, 2, \dots, K-1\}$  if we assume that  $p_k \rightarrow 0$  as  $k \rightarrow \infty$ . In our examples, we let  $K = 100$ . For example, to evaluate the PGF for cascade size, we iterate through eqs. (4.3) to (4.7) setting  $x = e^{-\frac{2\pi i}{K} l}$  and do this for each  $l$ , this gives us a vector holding values of the approximation of the discrete Fourier transform of the cascade size probabilities. Once we have evaluated the PGF at each of these points, we use the inverse fast Fourier transform on that set of transformed points to obtain an approximation for the first  $K$  probabilities from the PGF, the larger a value for  $K$  that we choose, the closer the approximation will be to the true values. In Fig. 2(a), we show the theoretical distribution of cascade size which is found using this method to invert the PGF for cascade size that we described in Section 4.1.

This method naturally extends to two dimensions (and beyond), allowing us to retrieve, for example, the joint probability distribution of cascade size and cumulative depth from Section 4.3. The joint probability distribution of random variables  $X_1$  and  $X_2$  is given by the set of probabilities  $\{p_{k_1, k_2}\}$  where  $p_{k_1, k_2}$  is the probability that  $X_1 = k_1$  and  $X_2 = k_2$ . The general joint PGF has the form

$$A(x, y) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} p_{k_1, k_2} x^{k_1} y^{k_2}. \quad (5.3)$$

and similar to the univariate case, if we evaluate the PGF for all combinations of  $x \in \{1, e^{-2\pi i/K_1}, e^{-4\pi i/K_1}, \dots, e^{-2(K_1-1)\pi i/K_1}\}$  and  $y \in \{1, e^{-2\pi i/K_2}, e^{-4\pi i/K_2}, \dots, e^{-2(K_2-1)\pi i/K_2}\}$ , we get the two-dimensional discrete Fourier transform of the joint probabilities of all combinations of the first  $K_1$  values for  $X_1$  and the first  $K_2$  values for  $X_2$ ; that is,  $k_1$  varying from 0 to  $K_1 - 1$  and  $k_2$  varying from 0 to  $K_2 - 1$ ;

$$P_{l,m} = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} p_{k_1, k_2} \left( e^{-\frac{2\pi i l}{K_1}} \right)^{k_1} \left( e^{-\frac{2\pi i m}{K_2}} \right)^{k_2} \approx A \left( e^{-\frac{2\pi i l}{K_1}}, e^{-\frac{2\pi i m}{K_2}} \right). \quad (5.4)$$

for  $K_1$  and  $K_2$  sufficiently large. In the case of the joint PGF for cascade size and cumulative depth, this means iterating through eqs. (4.21) to (4.24) with  $x = e^{-2\pi i l/K_1}$  and  $y = e^{-2\pi i m/K_2}$  for all  $l, m$  combinations,

$l \in \{0, 1, \dots, K_1 - 1\}$  and  $m \in \{0, 1, \dots, K_2 - 1\}$ . From eq. (5.4) if we evaluate the PGF for all combinations of  $x \in \{1, e^{-2\pi i/K_1}, e^{-4\pi i/K_1}, \dots, e^{-2(K_1-1)\pi i/K_1}\}$  and  $y \in \{1, e^{-2\pi i/K_2}, e^{-4\pi i/K_2}, \dots, e^{-2(K_2-1)\pi i/K_2}\}$  and take the two-dimensional inverse fast Fourier transform, we will recover an approximation for the joint probabilities  $p_{k_1, k_2}$  for all combinations of the first  $K_1$  values for  $k_1$  and the first  $K_2$  values for  $k_2$ . As for the uni-variate case, the accuracy in the approximation increases as  $K_1$  and  $K_2$  increase. It is worth noting that when we increase the number of dimensions, to keep the accuracy the same in the marginal distributions of each of the variables results in an exponential increase in the number of times that the PGF needs to be evaluated; for example, if we want to keep the accuracy of the marginal distribution when  $K = 100$  when we extend to five dimensions we need to evaluate the PGF  $10^5$  times. However, for most of the distributions that we have looked at where the probability  $p_k$  decays quickly as  $k \rightarrow \infty$  meaning that choosing  $K = 100$  gives us more accuracy than we need. For these types of distributions, it is reasonable to decrease the value of  $K$  if the computational power required in increasing the dimension becomes an issue. It is also possible for  $K_1, K_2, \dots, K_5$  to all have different values. If we use this method to recover the joint distribution for cascade size and cumulative depth from the joint PGF derived in Section 4.3, for example, we can calculate theoretical properties of the cascades including, but not limited to, the EATD and the correlation between cascade size and cumulative depth that we will describe in Section 5.1.

In  $N$  dimensions the PGF would take the form,

$$F(x_1, x_2, \dots, x_N) = \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \dots \sum_{k_N=0}^{\infty} p_{k_1, k_2, \dots, k_N} x_1^{k_1} x_2^{k_2} \dots x_N^{k_N}, \quad (5.5)$$

which, when evaluated for each combination of  $x_i \in \{1, e^{-2\pi i/K_i}, e^{-4\pi i/K_i}, \dots, e^{-2(K_i-1)\pi i/K_i}\}$ ,  $i \in \{1, 2, \dots, N\}$  becomes approximately the  $N$ -dimensional discrete Fourier transform of the probabilities  $p_{k_1, k_2, \dots, k_N}$ . As for the univariate and bivariate cases, taking the inverse fast Fourier transform gives us an approximation to the  $N$ -dimensional joint distribution.

### 5.1 Pearson correlation coefficient and EATD

In this subsection, we show how the Pearson correlation coefficient between cascade size and cumulative depth and the EATD are calculated from the joint probability distribution that we described in Section 4.3. The joint distribution is recovered from its PGF in the way that we have just shown. The Pearson correlation coefficient can be expressed in terms of the moments of the distribution

$$\rho_{\tilde{X}, \tilde{Y}} = \frac{\mathbb{E}[\tilde{X}\tilde{Y}] - \mathbb{E}[\tilde{X}]\mathbb{E}[\tilde{Y}]}{\sqrt{\mathbb{E}[\tilde{X}^2] - \mathbb{E}[\tilde{X}]^2} \sqrt{\mathbb{E}[\tilde{Y}^2] - \mathbb{E}[\tilde{Y}]^2}} \quad (5.6)$$

these moments can be calculated from the probability distribution. For unclustered networks, Gleeson *et al.* [6] described a method for calculating the EATD using an integral formula; however, their method does not extend to clustered networks. An alternative method of calculating the EATD is to use the joint probability distribution of cascade size and cumulative depth, this method is not restricted to unclustered networks. We compute the EATD using the formula

$$\text{EATD} = \sum_{i, j=0}^{\infty} P[\text{cascade size} = i \ \& \ \text{cumulative depth} = j] \frac{j}{i}. \quad (5.7)$$

TABLE 1 Table showing both the EATD and the Pearson correlation coefficient,  $\rho$ , from 1 million Monte-Carlo simulations and from the MTBP theory. We show results for a doubly Poisson Newman–Miller network ( $\mu = 1$ ,  $\nu = 4$ ) with 10,000 nodes and for two empirical networks; the power-grid network [3], the largest connected component of the network-science co-authorship network [36] and the *C. elegans* metabolic network [37]. For the empirical networks, in the calculation of the theoretical EATD and  $\rho$ , we estimated the Newman–Miller distribution ( $\pi_{st}$  from Section 3) using the edge-disjoint edge clique cover (EECC) described in Appendix B. The parameters are chosen such that the dynamics in the simulations on the empirical networks are characteristically close enough to criticality that we see a wide range of cascade sizes.

Network	$p_1$	$\alpha$	simulation: EATD	theory: EATD	simulation: $\rho$	theory: $\rho$
Newman–Miller	0.05	0	0.347	0.333	0.905	0.898
Newman–Miller	0.02	0.2	0.126	0.124	0.925	0.932
Power grid	0.04	0.15	0.062	0.059	0.796	0.863
Science co-authorship	0.01	0.1	0.041	0.027	0.938	0.960
<i>C. elegans</i>	0.005	0.005	0.007	0.026	0.976	0.935

In Table 1, we show the EATD and Pearson’s correlation found using this method and compare to Monte-Carlo simulations on two synthetic networks and three real-world networks.

## 6. Application to real-world networks

To apply these methods to empirical networks, we need to recover the joint distribution of triangles and single links from the network, this allows us to find the PGF for the distribution  $\tilde{f}(x, y)$  and for the excess distributions  $f_r(x, y)$  and  $f_q(x, y)$ , as given in Section 3, for a Newman–Miller network with the same joint distribution as the empirical network. An assumption of the MTBP method for modelling complex contagion on clustered networks is that the cliques do not share edges—they are edge-disjoint or, equivalently, are connected in a locally tree-like manner—in practice, many real-world networks have cliques which are not edge-disjoint. Burgio *et al.* [34] propose the Edge-Disjoint Edge Clique Cover (EECC), a cover which imposes, as the name suggests, the constraint that the cliques are edge-disjoint. With the EECC, we can approximate  $\tilde{f}$ , and thus  $f_q$  and  $f_r$ .

In our examples here, we assume that the network is composed of triangles and links but not higher-order cliques. Since the cliques in real-world networks can have more than three nodes; that is, a triangle, we approximate any higher-order cliques as a combination of triangles and links, as shown in Fig. 3(b). Fig. 3(a) shows the decomposition of the network into edge-disjoint cliques where there is no restriction on the size of the cliques, this is the EECC introduced by Burgio *et al.* [34]. In Appendix B, we present our alteration of the EECC algorithm introduced by Burgio *et al.* [34], which imposes that the cliques in the EECC have maximum size 3; that is, are at most triangles, this is the algorithm that we use here. This decision was due to the fact that the equations that we set up in previous sections assume that the network is fully composed of triangles and single links; in theory, if these systems of equations are extended to network distributions with higher-order cliques then it would be natural to extend the algorithm to capture these larger cliques. However, it is possible that extending the systems of equations beyond relatively small clique sizes would be prohibitively tedious by hand and an extension to higher order cliques is likely to require the derivation of the equations to be automated.

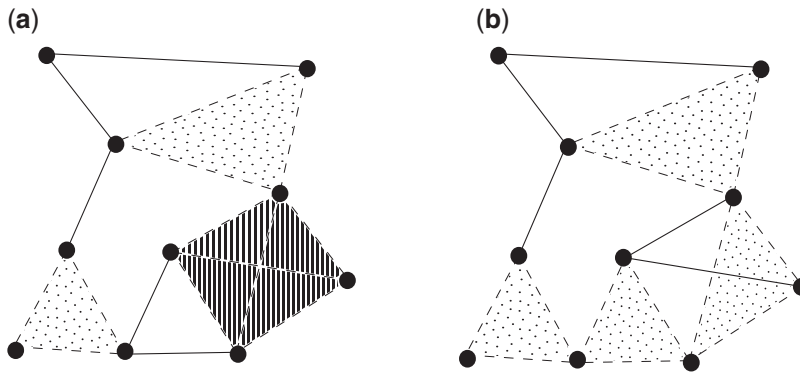


FIG. 3. (a) An edge-disjoint edge clique cover (EECC) of a network (black nodes and edges) showing the cliques included in the cover, the 3-cliques (dots) and 4-clique (stripes) are included in the EECC, the edges that do not form part of these cliques are included in the EECC as 2-cliques. The dashed edges represent edges in a higher-order clique and the solid edges connect two nodes in a 2-clique. (b) The EECC of the same network when we restrict the maximum clique size to 3.

### 6.1 Implementation of the EECC

In this subsection, we apply the EECC to a synthetic network generated from the doubly Poisson distribution and compare the results to what would be attained from a single-edge decomposition (SED), which assumes that the network is fully composed of single edges and thus the spreading dynamics constitute a simple branching process. In Fig. 4, we show the results of using the EECC on a synthetically generated network with a doubly Poisson distribution with parameters  $\mu = 1$  and  $\nu = 4$  (mean degree 9). In this network, the assumptions of the MTBP theory are very close to being met; that is, the cliques do not overlap and the network is very large (5000 nodes). We applied the EECC to the network to approximate its joint triangle-link distribution and use this to find the distribution of cascade sizes according to the method described in Section 4.1. We compare the results of this approach to using the SED and the method for calculating the cascade size distribution given in [6], for completeness, this method is summarized in Appendix A.

In Fig. 4(a, b) we examine the distribution of cascade sizes for simple-contagion dynamics where  $p_1 = 0.05$  and  $\alpha = 0$ . For a simple contagion on this network, we see that the theoretical distribution of cascade size matches well to the simulations when we account for the presence of triangles in the network. However, when we compare the MTBP which accounts for the presence of triangles to a simple branching process approximation (using the SED) which does not, there is minimal difference in the probability distributions, both curves match the simulations very well—this result is not particularly surprising. Melnik *et al.* [7] found that tree-based theory for dynamical processes on various highly clustered networks yielded accurate results despite the locally tree-like assumption being clearly violated. In Fig. 4(c) and (d), we examine the cascade size distribution for complex-contagion dynamics where  $p_1 = 0.02$  and  $\alpha = 0.2$ , we compare simulations on the network to the theoretical distribution of sizes and find very strong agreement between the theory and simulations. In the figures, we include the simple branching process approximation of the same dynamics and see that accounting for the clustering in the network gives a more accurate theoretical result. When we compare the result for the theory that accounts for the presence of triangles to the theory that does not, we see that for these parameters not accounting for the presence of triangles leads to smaller cascades than are observed in simulations, again showing the importance of accounting for clustering when modelling complex contagion on clustered networks.

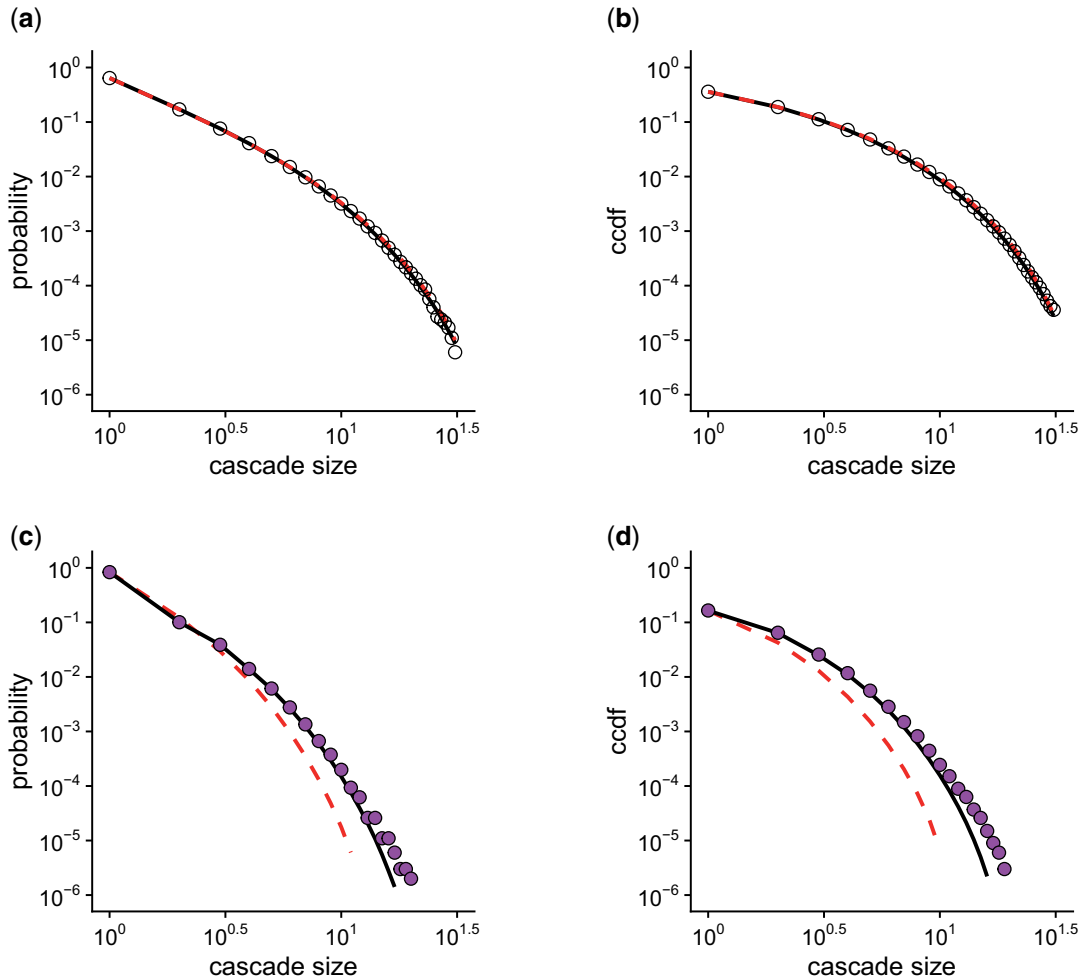


FIG. 4. The PMFs (probability distributions) (a) and (c) and CCDFs (b) and (d) for cascade size for a synthetically generated network of size 5,000 which follows a doubly Poisson distribution with parameters  $\mu = 1$  and  $\nu = 4$  for the dynamics described in Section 2. In (a) and (b), the complex contagion parameters are  $p_1 = 0.05$  and  $\alpha = 0$ , this is a simple contagion, in (c) and (d), the parameters are  $p_1 = 0.02$  and  $\alpha = 0.2$ , this is a complex contagion. The points are from simulating 1 million cascades on the network, the solid line is from estimating the PGF of the link-triangle distribution from the network using an EECC and the dashed line is a simple branching process approximation of the dynamics. In (a) and (b), we have removed the colour from the points to make the curves from theory more visible.

## 6.2 Application of the EECC to real-world networks

So far, we have shown that we can derive interesting statistics about complex-contagion dynamics on clustered networks—or at least networks which are structurally composed of cliques which are edge disjoint. While we remain in the (arguably unrealistic) regime where the maximum clique size is 3, we would like to know how these methods work when we apply them to real-world networks which may have correlations and clique size distributions which are not present in the synthetic distributions that we have worked with so far. We apply these methods to three well-studied real-world networks; the power-grid

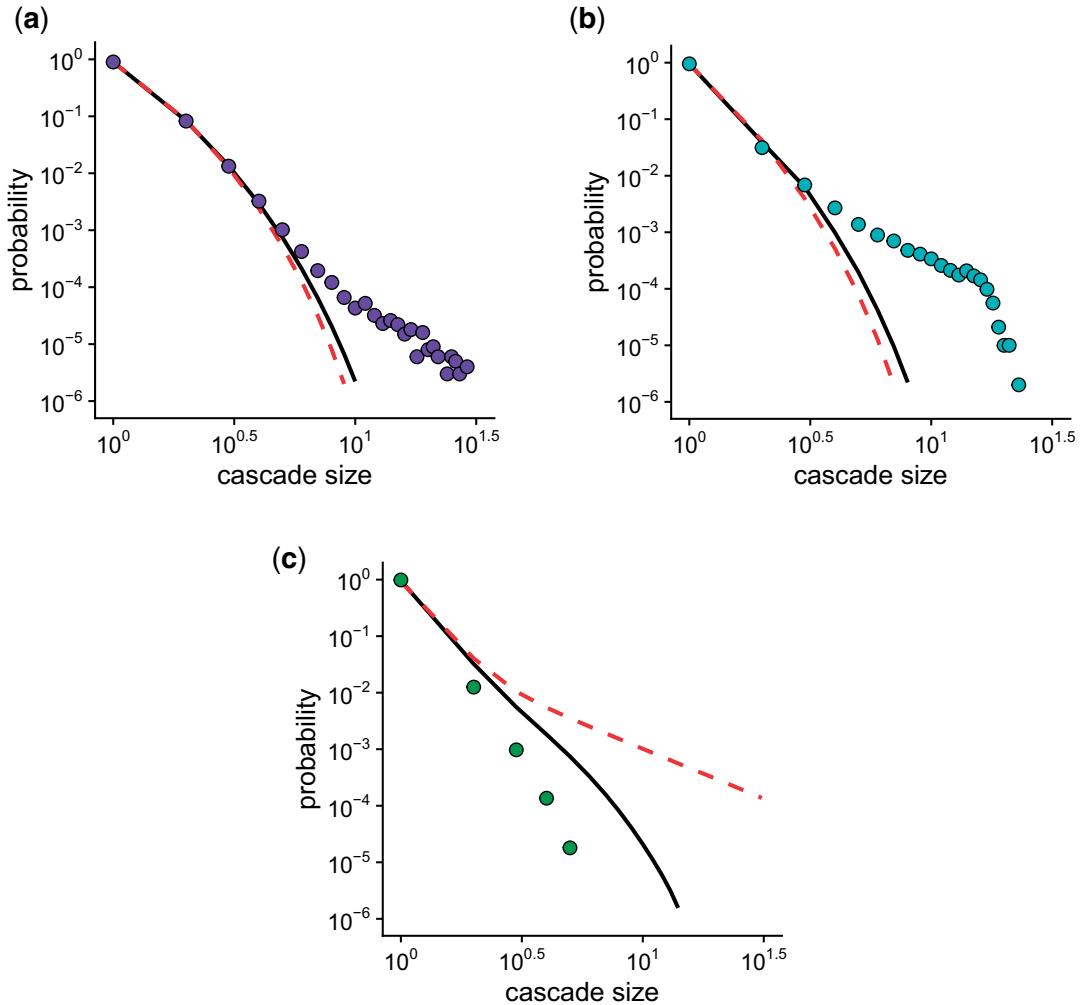


FIG. 5. The probability distributions for cascade size for complex-contagion dynamics on (a) the power-grid network [3] ( $p_1 = 0.04$  and  $\alpha = 0.15$ ), (b) the largest connected component of the science co-authorship network [36] ( $p_1 = 0.01$  and  $\alpha = 0.1$ ) and (c) the *C. elegans* metabolic network [37] ( $p_1 = 0.005$  and  $\alpha = 0.005$ ). In all subfigures, we compare the theoretical distribution for cascade size accounting for the presence of triangles (network distribution approximated using the EECC) (solid line) and not accounting for the presence of triangles (dashed line) to 1 million simulations (points).

network [3] (4941 nodes, 6594 edges,  $C_\Delta = 0.10^1$ ), the largest connected component of the science co-authorship network [36] (379 nodes, 914 edges,  $C_\Delta = 0.43$ ) and the *C. elegans* metabolic network [37] (453 nodes, 4596 edges,  $C_\Delta = 0.12$ ). We chose these networks because they are relatively large—reducing the impact of finite-size effects—but small enough that the EECC algorithm finds a cover in a reasonable run time. In Figs. 5(a) and 6(a), we show the probability distribution and CCDF, respectively

<sup>1</sup> The clustering coefficient,  $C_\Delta$ , is the ratio between the number of triangles and the number of connected triples in the network.

of cascade size for the power-grid network with complex contagion parameters  $p_1 = 0.04$  and  $\alpha = 0.15$ . We compared 1 million simulations on the network to the theoretical distributions accounting for triangles with the MTBP and also not accounting for triangles using the SED. While the theory matches quite well to the simulated cascade size probability distribution, in Figs. 5(a) and 6(a), for smaller cascade sizes, the difference in what is predicted by the SED and by the MTBP with the EECC is very small. The tree-like approximation made in the SED does quite well despite the power-grid network being clustered. In the tails of the cascade size distribution both of the theoretical distributions do not match the simulations very well, we speculate that this is due to the larger cliques present in the network which are not accounted for in the theoretical approach. Similarly, we applied the methods to the largest connected component of the science co-authorship network [36], using complex contagion parameters  $p_1 = 0.01$  and  $\alpha = 0.1$ , we show the probability distribution and CCDF in Figs. 5(b) and 6(b). The EECC was also applied to this network in Ref. [38]. We found that while the simulations match the theory extremely well for cascade sizes up to size three, which accounts for 99.2% of all cascades, our theory with the EECC under-predicts the number of large cascades that we observe in simulations. We anticipate that this is because there are larger cliques in the empirical network that are not accounted for when we use the EECC with maximum clique size 3 to approximate the clique distribution of the network. Given that this network is not very large (379 nodes) and that branching processes assume an infinitely large network, it is important to check that the network size is not causing the discrepancy between the theoretical distributions and the simulated cascades. In Appendix C, we explore the impact of finite-size effects and find that these are minimal, again reinforcing our hypothesis that the difference between the theory and the simulations is caused by the theory not accounting for the higher-order cliques (with more than three nodes). Finally, we applied this method to the *C. elegans* metabolic network [37] for complex contagion with parameters  $p_1 = 0.005$  and  $\alpha = 0.005$ . For this network, the simple branching process theory poorly approximates the cascade-size distribution (probability distribution and CCDF) as shown by the red dashed line in Figs. 5(c) and 6(c), respectively. When we account for the presence of triangles (as shown by the black solid line) using the MTBP with the EECC to approximate the Newman–Miller distribution, the theory comes much closer to the simulations on the empirical network which suggests to us that it is the presence of higher-order structures in the *C. elegans* metabolic network that leads the simple branching process to poorly describe the dynamics. By accounting for links and triangles only, the MTBP substantially improves the theoretical distribution of cascade size, when we compare to simulations on the *C. elegans* metabolic network.

## 7. Discussion and Conclusions

In this article, we have extended the multi-type branching process (MTBP) theory originally proposed in Ref. [16] in three main directions. First, the clustered networks studied in [16] were highly stylized, and we showed how the more general Newman–Miller class of networks can be used to incorporate clustering into the branching process framework; then, we used the MTBP to derive more results, the focus here was on looking at full distributions of cascade sizes, cascade lifetimes and the joint distribution of cascade size and cumulative depth which enabled us to calculate the expected average tree depth and the correlation between cascade size and cumulative depth, this contrasts with a focus on the average behaviour of the cascades in our previous article [16]. Finally, we showed how to apply the theory to real-world networks. To apply the theory to real-world networks, we approximated the real-world network structure as a Newman–Miller network using the edge-disjoint edge clique cover (EECC) proposed by Burgio *et al.* [34]. A second contribution of this article is our derivation of the numerical inversion method for multi-variate probability generating functions (PGFs). While Cavers’ method [31] has been extended

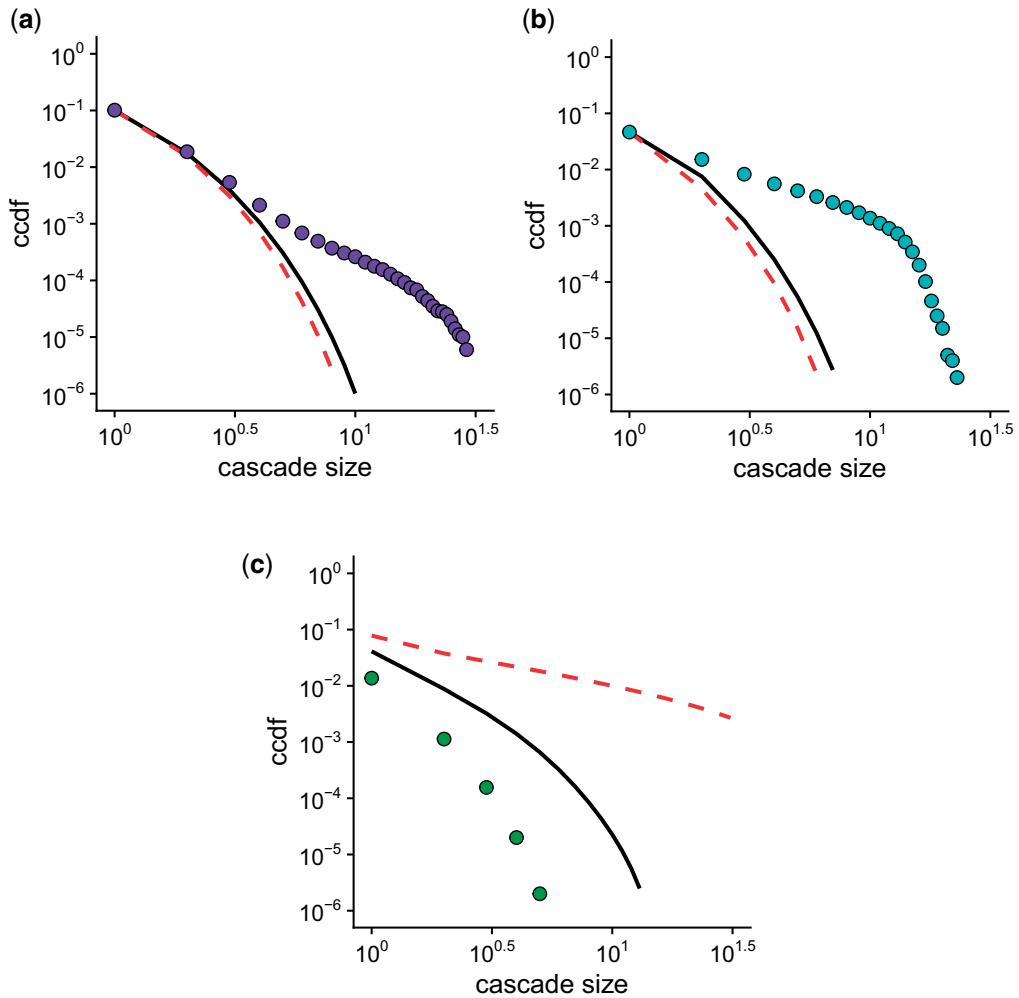


FIG. 6. The complimentary cumulative distribution function (CCDF) for cascade size for complex-contagion dynamics on (a) the power-grid network [3] ( $p_1 = 0.04$  and  $\alpha = 0.15$ ), (b) the largest connected component of the science co-authorship network [36] ( $p_1 = 0.01$  and  $\alpha = 0.1$ ) and (c) the *C. elegans* metabolic network [37] ( $p_1 = 0.005$  and  $\alpha = 0.005$ ). In all subfigures, we compare the theoretical distribution for cascade size accounting for the presence of triangles (network distribution approximated using the EECC) (solid line) and not accounting for the presence of triangles (dashed line) to 1 million simulations (points).

to bivariate PGFs [33], our alternative derivation has the advantage of extending naturally to PGFs of any dimension.

In Keating *et al.* [16], we focused on a very stylized model of network clustering for ease of explanation. These networks were infinitely large, and we assumed that the local structure of the network was invariant throughout the network; for example, we looked at networks where every node was in two 4-cliques and a network where every node was in four 2-cliques and one 3-clique. Here, we showed how the method can be extended to incorporate any Newman–Miller network distribution describing the distribution of triangle and single-link membership of the nodes in a network. In our examples, we use

the doubly Poisson distribution as used by Newman [29]. While we did not delve into it here, using the probability generating functions for the Newman–Miller network distributions, all of the results on the average behaviour of the cascades from Ref. [16] can be extended to this more general class of networks. For example, we can calculate the expectations from the offspring distribution PGFs and use this to find the elements of the mean matrix, allowing us to find the cascade condition and the expected cascade size using the methods described in Ref. [16]. Furthermore, while we focused here on Newman–Miller networks, which have a maximum clique size of three, in principle the method can be extended beyond this restriction to multivariate distributions of cliques of any size, this would mean that we would need to incorporate different *types* into the MTBP for which the transition probabilities may be calculated either by hand or using a computer algorithm. To incorporate very large cliques into the MTBP, we expect that it will not be feasible to do this by hand and it would be necessary to write a computer algorithm to do this for us—we did not attempt to do this as part of this work.

Gleeson *et al.* [6] successfully used simple branching processes to model information cascades on Twitter. They analytically derived full distributions of cascade lifetimes and cascade size, and calculated the expected average tree depth (EATD) and structural virality. For complex-contagion dynamics on clustered networks, we here showed how, using PGFs, we can find the cascade size distribution, distribution of cascade lifetimes and the EATD. Our method for calculating the cascade size distribution, as in Ref. [6], involves finding the PGF for the cascade size distribution and recovering the probability distribution using an inverse fast Fourier transform method. In Section 5, we explained how inverse fast Fourier transforms can be used to recover an accurate approximation of the probabilities, this naturally extends to higher-dimensional PGFs. We use inverse fast Fourier transforms to invert a two-dimensional PGF for the joint distribution of cascade size and cumulative depth, allowing us to find the EATD. The inverse fast Fourier transform method is crucial for finding accurate distributions under the MTBP theory, other studies use computer algebra systems [32], which are limited by the number of terms that can be computed in the probability distribution and by the number of generations that can be accounted for in the branching process. The inability for computer algebra systems to recover cascade-size probabilities for branching processes that are not truncated after a short number of generations means that they can not accurately model many real-world processes that might survive beyond a small number of generations.

In Section 6 of this article, we applied the MTBP method to synthetic and real-world networks. Using the edge-disjoint edge clique cover (EECC) method from Burgio *et al.* [34], we recovered the Newman–Miller distribution of both a synthetically generated network (Fig. 4) and a Newman–Miller approximation for the clique distribution of three real-world networks (Figs. 5 and 6) and compared the cascade-size distribution from the MTBP theory for given complex-contagion dynamics to a simulation of the dynamics on the empirical network. For the synthetically generated network, the match between the theory and simulations is very close with only small finite-size effects—in MTBPs we assume that the network is infinitely large. The MTBP theory better models the dynamics for complex contagion, but for simple-contagion dynamics, there is very little difference in the accuracy of the theory using simple branching processes in comparison to MTBPs, the simple branching process is very accurate despite the network violating the locally tree-like assumption. For the real-world networks, we also get a good match between the theory and simulations for small cascade sizes (cascades up to size 3 account for over 99% of cascades); however, for larger cascade sizes the MTBP is less accurate, this is likely because the clique cover decomposes larger cliques into a combination of triangles and single links. If very large cliques are present in the network when there is social reinforcement present; that is,  $\alpha > 0$ , the simulated cascades can be much larger than those predicted by the theory. In all real-world networks that we show here, the MTBP is more accurate in predicting the cascade size distribution than a simple branching process

approximation; however, an issue with using real-world networks is that if they are too small we often encounter finite-size effects (which we explore further in Appendix C) and if they are too large the EECC algorithm can take too long to converge.

The examples here were restricted to Newman–Miller networks, where we assume that the cliques never have more than three nodes; however, in many real-world networks, this is an over-simplification and not accounting for larger cliques can lead to less accurate results. If future models were to account for these larger cliques, we would expect more accurate results. We have not extended the model beyond cliques of size three as deriving the PGFs for quantities such as cascade size by hand to include very large clique sizes would take a very long time. We imagine that it would be possible to write a computer algorithm to derive the equations; however, this is beyond the scope of this article. We also assumed that the cliques do not share edges; that is, are edge-disjoint, in reality this is not always the case; however, it is possible to extend the model to account for this by including motif types which are arrangements of three cliques which share edges.

### A. Simple branching process approximation for the cascade size calculation

For comparison with the results that account for triangles, here we show how to approximate the dynamics with a single-edge decomposition (SED); that is, assuming that the network is locally tree-like. As in Section 4.1, let  $\tilde{X}_n$  be the size of a full cascade allowed to grow for  $n$  generations; then, we can write  $\tilde{X}_n$  in terms of the sizes of its subtrees,

$$\tilde{X}_n = 1 + \sum_i X_n^{(i)}, \quad (\text{A.1})$$

where  $X_n^{(i)}$  is the size of subtree  $i$  allowed to grow for  $n$  generations. Similar to in Section 4.1, the size of a subtree does not include its seed node, in the SED there is only one subtree type and it is the same as a type-5 subtree in Section 4.1. The PGF for cascade size is

$$\begin{aligned} \tilde{K}_n(z) &= \mathbb{E} \left[ z^{\tilde{X}_n} \right] \\ &= \sum_m \pi_m \mathbb{E} \left[ z^{1 + \sum_{i=1}^m X_n^{(i)}} \right] \\ &= \sum_m \pi_m z \mathbb{E} \left[ z^{X_n} \right]^m \\ &= z \tilde{f}(K_n(z)), \end{aligned} \quad (\text{A.2})$$

where  $\pi_m$  is the probability that a node chosen at random has degree  $m$  and  $\tilde{f}(z)$  is the PGF for the degree distribution. Similarly,

$$\begin{aligned} K_n(z) &= \mathbb{E} \left[ z^{X_n} \right] \\ &= 1 - p_1 + p_1 \mathbb{E} \left[ z^{1 + \sum_i X_{n-1}^{(i)}} \right] \\ &= 1 - p_1 + p_1 z \sum_m q_m \mathbb{E} \left[ z^{X_{n-1}} \right]^m \\ &= 1 - p_1 + p_1 z f(K_{n-1}(z)), \end{aligned} \quad (\text{A.3})$$

where  $q_m$  is the probability that the excess degree of a node reached by traversing a link is  $m$  and  $f(z)$  is the PGF for the excess degree distribution. To find the full PGF at the point  $z$ , we iterate through equations eq. (A.2) and eq. (A.3) with initial conditions  $K_0(z) = 1$  until  $K_n(z)$  is within a tolerance of  $10^{-5}$  of  $K_{n-1}(z)$  when evaluated for a specific value for  $z$ . We then find the full probability distribution from the PGF using the method described in Section 5.

### B. EECC algorithm with maximum clique size 3

In this section, we detail the EECC algorithm that we use in Section 6, this algorithm is a modification of the EECC introduced by [34] which was suggested but not explicitly presented in the article [34]. In our modification, we restrict the clique size to a maximum of 3, naturally this can be extended to larger maximum sizes. The steps are as follows:

1. Let  $G$  be the network, find  $C$ , the set of maximal cliques in  $G$ .
2. Find all of the elements of  $C$  which have order not greater than 3 and do not share edges with other elements of  $C$  and add these elements (cliques) to the EECC.
3. Remove any elements from  $C$  that are in the EECC.
4. Let  $C_{OL}$  be the set of all cliques in  $C$  that share at least one edge with another clique in  $C$  and let  $C_{HO}$  be the set of all elements of  $C$  that are of order greater than 3.
5. Remove any elements from  $C_{OL}$  that are also in  $C_{HO}$ .
6. While  $C_{OL}$  is not empty, for each element calculate  $\rho$ , the proportion of edges shared with other elements of  $C_{OL}$ . Add a randomly chosen element with the lowest  $\rho$  to the EECC and remove from  $C_{OL}$ . Remove all edges present in the EECC from  $C_{OL}$ . Recalculate the maximal cliques and  $\rho$  considering all remaining elements.
7. While  $C_{HO}$  is not empty, for each element calculate  $\rho$ , the proportion of edges shared with other elements of  $C_{HO}$ . For a randomly chosen element with the smallest value for  $\rho$ , if it is of order not greater than 3 add to the EECC; otherwise, add a subclique of size 3 to the EECC. Remove all edges present in the EECC from  $C_{HO}$ , recalculate the maximal cliques and  $\rho$  considering all remaining elements.

### C. Finite-size effects

In Figs. 5 and 6, we see differences between the cascade-size distribution for the Monte-Carlo simulations on the empirical networks and the cascade-size distributions from the MTBP theory, especially in the tails of the distributions. One potential cause for this is that, since the networks are not very large (the science co-authorship network has 379 nodes), these could be finite-size effects. When we model cascades on networks using branching processes, we make the assumption that the network is infinitely large; however, this assumption is clearly violated in the case of these real-world networks. Another possible cause of this differences between the theory and the simulations is that the real-world networks contain higher-order cliques and highly connected structures that are lost in the MTBP model when we approximate the network structure using the EECC with maximum clique size of three.

All cliques with more than three nodes are decomposed into a series of edge-disjoint triangles and single links. To determine whether these discrepancies are predominantly due to finite-size effects, we generate a synthetic network of approximately the same size as the empirical networks from the Newman–Miller distributions calculated from their EECC and run 1 million Monte-Carlo simulations on each

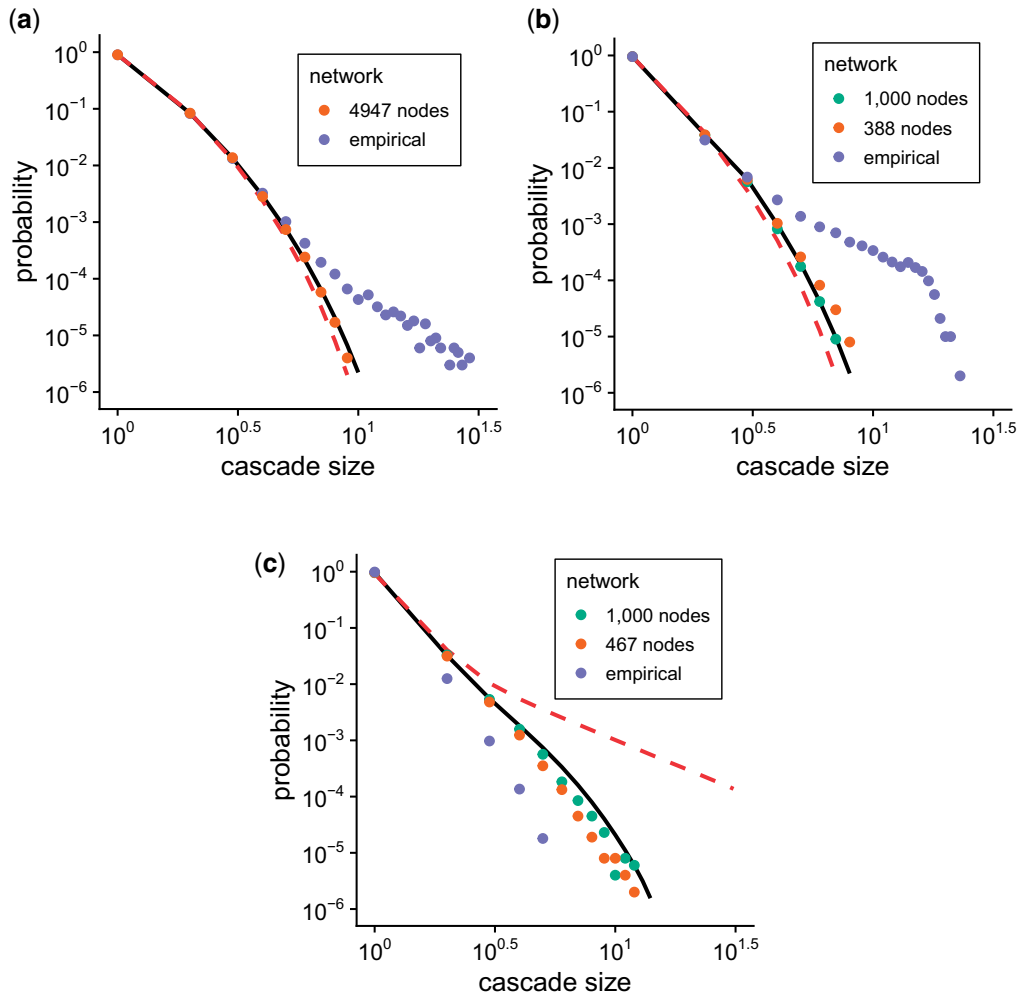


FIG. C.1. The aim of this figure is to explore the impact of finite-size effects in the Monte-Carlo simulations. For each of the real-world networks studied, we constructed a synthetic Newman–Miller network of approximately the same size as the empirical network using the Newman–Miller distribution retrieved from finding an EECC on the empirical network. We performed 1 million Monte-Carlo simulations on each of these networks. For each of these networks, the cascade-size distribution on the empirical network is shown by the purple points and the simulations on the synthetic network of approximately the same size is shown by the orange points. The closeness between the orange points and the theory suggests that the impact of finite-size effects is minimal in causing the difference between the theory and simulations on the empirical network. In all subfigures, we compare the theoretical distribution for cascade size accounting for the presence of triangles (network distribution approximated using the EECC) (black solid line) and not accounting for the presence of triangles (dashed line) to 1 million simulations (points). (a) powergrid: probability distribution. (b) science co-authorship: probability distribution. (c) *C. elegans*: probability distribution.

of these synthetic networks. These synthetic networks meet the assumption that is made in the theory that the network is fully composed of triangles and single links. Therefore, we can disentangle the impact of finite-size effects from other structural features present in the empirical network which are not within the assumptions of the MTBP theory. In Fig. C.1, we show the cascade-size distributions for the

powergrid network [3], the largest connected component of the science co-authorship network [36] and the *C. elegans* network [37] for both the MTBP theory (black solid line), the simulations on the empirical network (purple points) and simulations on a synthetic network for a similar size generated from the Newman–Miller distribution obtained from the EECC on that network (orange points).

We find that the differences between the theory and the simulations on the synthetic network are minimal in comparison to the differences between the theory and the simulations on empirical networks. This confirms that the impact of finite-size effects is minimal for the simulations and that the discrepancies between the distributions from the MTBP theory and the Monte-Carlo simulations on real-world networks are most likely due to the fact that there are differences other than size in the network structure between the empirical network and in the distribution that we use in the MTBP from the EECC.

## Funding

Science Foundation Ireland [18/CRT/6049 to L.K., 16/IA/4470 to J.G., 16/RC/3918 to J.G., 12/RC/2289 P2 (J.G.)] with co-funding from the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## REFERENCES

1. NEWMAN, M. E. & PARK, J. (2003) Why social networks are different from other types of networks. *Phys. Rev. E*, **68**, 036122.
2. PORTER, M. A. & GLEESON, J. P. (2016) *Dynamical Systems on Networks: A Tutorial*. Cham: Springer.
3. WATTS, D. J. & STROGATZ, S. H. (1998) Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440–442.
4. WENG, L., MENCZER, F. & AHN, Y. Y. (2013) Virality prediction and community structure in social networks.. *Sci. Rep.*, **3**, 1–6.
5. GLEESON, J. P. (2013) Binary-state dynamics on complex networks: pair approximation and beyond. *Phys. Rev. X*, **3**, 021004.
6. GLEESON, J. P., ONAGA, T., FENNELL, P., COTTER, J., BURKE, R. & O’SULLIVAN, D. J. (2021) Branching process descriptions of information cascades on Twitter. *J. Complex Netw.*, **8**, 1–29.
7. MELNIK, S., HACKETT, A., PORTER, M. A., MUCHA, P. J. & GLEESON, J. P. (2011) The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E*, **83**, 036112.
8. CENTOLA, D. (2010) The spread of behavior in an online social network experiment. *Science*, **329**, 1194–1197.
9. KARSAI, M., IÑIGUEZ, G., KASKI, K. & KERTÉSZ, J. (2014) Complex contagion process in spreading of online innovation. *J. R. Soc. Interface*, **11**, 20140694.
10. MONSTED, B., SAPIEŻYŃSKI, P., FERRARA, E. & LEHMANN, S. (2017) Evidence of complex contagion of information in social media: an experiment using Twitter bots. *PLoS One*, **12**, e0184148.
11. ROMERO, D. M., MEEDER, B. & KLEINBERG, J. (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, **1**, 695–704.
12. SPRAGUE, D. A. & HOUSE, T. (2017) Evidence for complex contagion models of social contagion from observational data. *PLoS One*, **12**, 40–42.
13. SHIRLEY, M. D. & RUSHTON, S. P. (2005) The impacts of network topology on disease spread. *Ecol. Complex*, **2**, 287–299.
14. O’SULLIVAN, D. J. P., O’KEEFFE, G. J., FENNELL, P. G. & GLEESON, J. P. (2015) Mathematical modeling of complex contagion on clustered networks. *Front. Phys.*, **3**, 71.
15. WATTS, D. J. (2002) A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. USA*, **99**, 5766–5771.

16. KEATING, L. A., GLEESON, J. P. & O’SULLIVAN, D. J. (2022) Multitype branching process method for modeling complex contagion on clustered networks. *Phys. Rev. E*, **105**, 0343306.
17. KEMPE, D., KLEINBERG, J. & TARDOS, E. (2003) Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **1**, 137–146.
18. ARAGÓN, P., GÓMEZ, V., GARCÍA, D. & KALTENBRUNNER, A. (2017) Generative models of online discussion threads: state of the art and research challenges. *J. Internet Serv. Appl.*, **8**, 15. <https://doi.org/10.1186/s13174-017-0066-z>.
19. GLEESON, J. P., CELLAI, D., PORTER, M. A. & REED-TSOCHAS, F. (2014) A simple generative model of collective online behavior. *Proc. Natl. Acad. Sci.*, **111**, 10411–10415.
20. GLEESON, J. P. & DURRETT, R. (2017) Temporal profiles of avalanches on networks. *Nat. Commun.*, **8**, 1–13.
21. GLEESON, J. P., O’SULLIVAN, K. P., BAÑOS, R. A. & MORENO, Y. (2016) Effects of network structure, competition and memory time on social spreading phenomena. *Phys. Rev. X*, **6**, 021019.
22. GLEESON, J. P., WARD, J. A., O’SULLIVAN, K. P. & LEE, W. T. (2014) Competition-induced criticality in a model of meme popularity. *Phys. Rev. Lett.*, **112**, 048701.
23. MCSWEENEY, J. K. (2020) Single-seed cascades on clustered networks. *Netw. Sci.*, **9**, 59–72.
24. VAZQUEZ, A. (2006) Spreading dynamics on heterogeneous populations: multitype network approach. *Phys. Rev. E*, **74**, 066114.
25. VAZQUEZ, A. (2021) Multitype branching and graph product theory of infectious disease outbreaks. *Phys. Rev. E*, **103**, L030301.
26. CASWELL, H. (2018) *Matrix Population Models*. Sunderland, MA, USA: Sinauer.
27. GOEL, S., ANDERSON, A., HOFMAN, J. & WATTS, D. J. (2016) The structural virality of online diffusion. *Manag. Sci.*, **62**, 180–196.
28. O’BRIEN, J. D., ALETA, A., MORENO, Y. & GLEESON, J. P. (2020) Quantifying uncertainty in a predictive model for popularity dynamics. *Phys. Rev. E*, **101**, 062311.
29. NEWMAN, M. E. (2009) Random graphs with clustering. *Phys. Rev. Lett.*, **103**, 058701.
30. MILLER, J. C. (2009) Percolation and epidemics in random clustered networks. *Phys. Rev. E*, **83**, 020901.
31. CAVERS, J. (1978) On the fast Fourier transform inversion of probability generating functions. *IMA J. Appl. Math.*, **22**, 275–282.
32. BRUMMITT, C. M., D’SOUZA, R. M. & LEICHT, E. (2012) Suppressing cascades of load in interdependent networks. *Proc. Nat. Acad. Sci. USA*, **109**, 12.
33. ANTAL, T. & KRAPIVSKY, P. (2010) Exact solution of a two-type branching process: clone size distribution in cell division kinetics. *J. Stat. Mech.*, **1**, P07028.
34. BURGIO, G., ARENAS, A., GÓMEZ, S. & MATAMALAS, J. T. (2021) Network clique cover approximation to analyze complex contagions through group interactions. *Commun Phys*, **4**, 111.
35. MANN, P., SMITH, V. A., MITCHELL, J. B. & DOBSON, S. (2022) Degree correlations in graphs with clique clustering. *Phys. Rev. E*, **105**, 044314.
36. NEWMAN, M. E. (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, **74**, 036104.
37. DUCH, J. & ARENAS, A. (2005) Community detection in complex networks using extremal optimization. *Phys. Rev. E*, **72**, 027104.
38. MANN, P. (2022) On the use of generating functions for topics in clustered networks. PhD Thesis, University of St. Andrews, St. Andrews, Fife, Scotland.