Cascade Size Distributions and Why They Matter

Rebekka Burkholz[1], **John Quackenbush**[1]
[1] Harvard T.H. Chan School of Public Health
655 Huntington Avenue
Boston, Massachussetts 02115
rburkholz@hsph.harvard.edu

Abstract

How likely is it that a few initial node activations are amplified to produce large response cascades that span a considerable part of an entire network? Our answer to this question relies on the Independent Cascade Model for weighted directed networks. In using this model, most of our insights have been derived from the study of average effects. Here, we shift the focus on the full probability distribution of the final cascade size. This shift allows us to explore both typical cascade outcomes and improbable but relevant extreme events. We present an efficient message passing algorithm to compute the final cascade size distribution and activation probabilities of nodes conditional on the final cascade size. Our approach is exact on trees but can be applied to any network topology. It approximates locally tree-like networks well and can lead to surprisingly good performance on more dense networks, as we show using real world data, including a miRNA-miRNA probabilistic interaction network for gastrointestinal cancer. We demonstrate the utility of our algorithms for clustering of nodes according to their functionality and influence maximization.

Introduction

The Independent Cascade Model (ICM) is a cornerstone in the study of spreading processes on networks. It has been proven useful in the source detection of epidemic outbreaks (Leskovec et al. 2007; Farajtabar et al. 2015; Xu and Chen 2015; Zhu, Chen, and Ying 2017), identification of fake news (Tschiatschek et al. 2018; Vosoughi, Roy, and Aral 2018), marketing (Leskovec, Adamic, and Huberman 2007; Kempe, Kleinberg, and Tardos 2005), or identification of causal miRNAs for cancer (Nalluri et al. 2017). Many related optimization algorithms require sampling from the model. A famous example is given by influence maximization (IM) (Kempe, Kleinberg, and Tardos 2003). As in many other applications, the main quantity of interest is the average final cascade size $\mathbb{E}(\rho)$. This is is to be maximized by selecting an appropriate set of seeds $\mathcal{S}$. The seeds play crucial roles in their respective networks. For example, they can be regarded as influential individuals in a social network or assumed to control signaling pathways in cancer. The choice of such seeds substantially influences the probability distribution of the final cascade. In analyzing networks, we find the distribution can be broad and multi-modal (Burkholz, Herrmann, and Schweitzer 2018; Burkholz 2019).

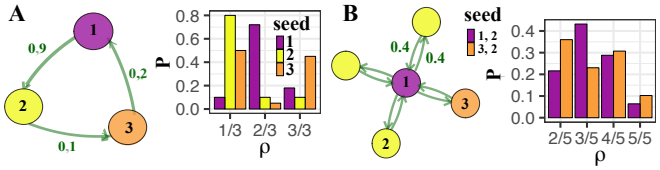

Figure 1: IM based on the average cascade size or the tail favors different initial seeds. The ICM is visualized on the left, the cascade size distributions corresponding to different initial seeds are shown on the right. Seed 1 (purple) maximizes the probability of large cascades, while Seed 3 (orange) maximizes the average cascade size. In B, Seed 2 is chosen in addition to either 1 or 3.

The average does not summarize such a distribution well, as it does not correspond to a probable event and so may not reflect the preference of decision makers or evolutionary pressure selecting seeds. Minimal examples, such as loops and stars, that tend to be common network motifs in larger real networks are shown in Fig. 1. While the purple Seed 1 maximizes the average cascade size, the orange Seed 3 maximizes the probability of larger cascades. In the examples provided, maximizing the probability of large cascades also has the effect of maximizing the probability of smaller ones. Risk averse decision makers might opt for more predictable outcomes and thus a more concentrated probability distribution.

Contributions

The shape of the cascade size distribution matters and our main contribution is the development of an efficient message passing algorithm for its computation: Subtree Distribution Propagation (SDP) is exact on trees and requires $O(N^2)$ computations. Parallelization can further speed-up computation. As extension to general networks, Tree Distribution Approximation (TDA), combines Belief Propagation (BP)

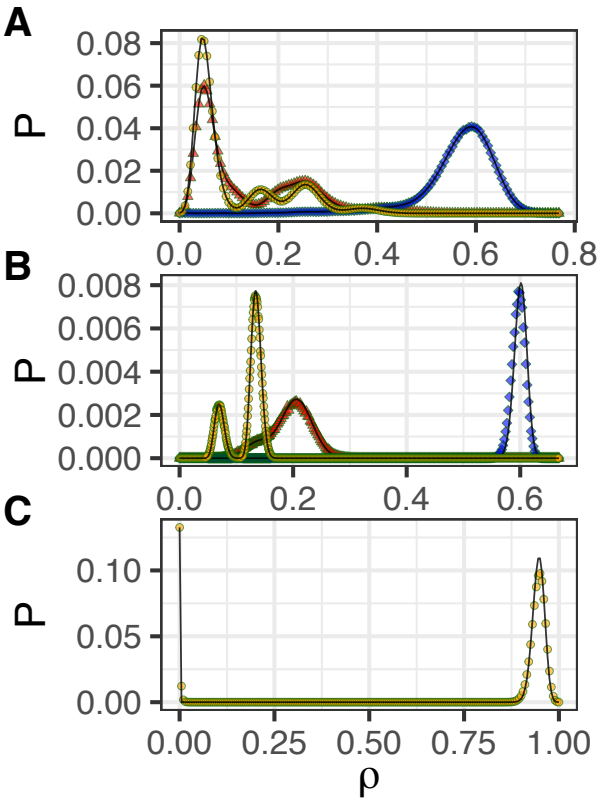

Figure 2: Final cascade size distribution for TDA (lines) and simulations (symbols). SD (orange circles), DD (blue squares), ED (red triangles). A. Artificial tree. B. Corporate ownership network. C. Cancer miRNA network with empirical weights.

with SDP. TDA is approximate, but accurate on locally tree-like networks. We can further compute marginal activation probabilities of nodes conditional on the final cascade size, a quantity no-one has studied before up to our knowledge. Yet, it provides rich information about the spreading behavior and susceptibility of nodes. We present two use cases: node clustering and influence maximization. 1) Conditional activation probabilities allow to identify nodes that are functionally similar for the cascade process. These similarities are usually caused by network symmetries. In a biological setting, these can imply redundant pathways and are particularly relevant to deepen our understanding of diseases like cancer. 2) Node similarities have also algorithmic implications. They reduce our set of seed candidates in influence maximization or related problems and thus speed up Greedy approaches. In some cases, they can even enable exhaustive search. Furthermore, we can use SDP or TDA to optimize different objectives than the average cascade size in influence maximization.

Alternatives like measures of the probability distribution tail are not submodular in general and, hence, do not provide a $(1 - 1/e - \epsilon)$ approximation guaranty for Greedy maximization as known for the original average cascade size. However, we show in experiments that an objective like expected shortfall can even achieve higher average cascade sizes in Greedy than the original approach. We study our algorithms using three data sets, an artificial tree to gain intuition, a larger real world corporate ownership network, and a dense network of miRNA signaling corresponding to gastrointestinal cancer.

Related Literature

As diverse as spreading phenomena are the related optimization objectives. The insight that the average cascade size is a submodular influence function (Kempe, Kleinberg, and Tardos 2003) has inspired many efforts to maximize this quantity by nearly optimal seed size selection (Du et al. 2014; Y. Lokhov and Saad 2016) or network adjustments (Wen et al. 2017). This has great applications, e.g., in marketing (Morris 2000; Goldenberg, Libai, and Muller 2001; Domingos and Richardson 2001). An alternative goal can be to balance information to avoid filter bubbles (Garimella et al. 2017) or overexposure (Abebe, Adamic, and Kleinberg 2017). Other works are more concerned with destructive aspects of cascades and their mitigation to avoid epidemic spreading (Budak, Agrawal, and El Abbadi 2011; Y. Lokhov and Saad 2016). Then, the (not always explicit) objective is to minimize cascades or to create boundary conditions that keep them small. Such analyses are usually based on average cascade sizes and/or sampling from the cascade model. To compute the average cascade size, faster alternatives to sampling are provided by local tree approximations (Newman 2002) for large random networks or belief propagation for smaller sparse networks (Gleeson and Porter 2018; Y. Lokhov and Saad 2016; Burkholz 2019). Yet, as recently shown (Burkholz, Herrmann, and Schweitzer 2018; Burkholz 2019), cascade size distributions can be broad and multi-modal so that the average does not provide a relevant summary statistic. Examples are provided in Fig. 2. The full probability distribution provides much richer information about a network structure. Usually, the extreme events are of highest interest to judge the robustness of a system or to find optima (Battiston et al. 2016; McNeil, Frey, and Embrechts 2015; Ohsaka and Yoshida 2017). For instance, (Ohsaka and Yoshida 2017) maximizes the expected shortfall of the final cascade size with respect to a portfolio of seeds by sampling from an ICM. Often, very large or small cascades occur only with small probabilities. This makes the optimization of tails harder, in particular, by sampling and demands alternatives. For a given simple ICM with uniform infection probability p (or a threshold model) and a locally tree-like network, the final cascade size distribution can be computed by message passing (Burkholz 2019). Our approach is inspired, even though the math is different and our algorithms are more efficient, as they need only a single Fourier Transformation. Furthermore, we capture general ICMs with heterogeneous weights and provide activation probabilities of nodes conditional on the final cascade size. In addition, we employ our algorithms to influence maximization and variants based on the full cascade size distribution.

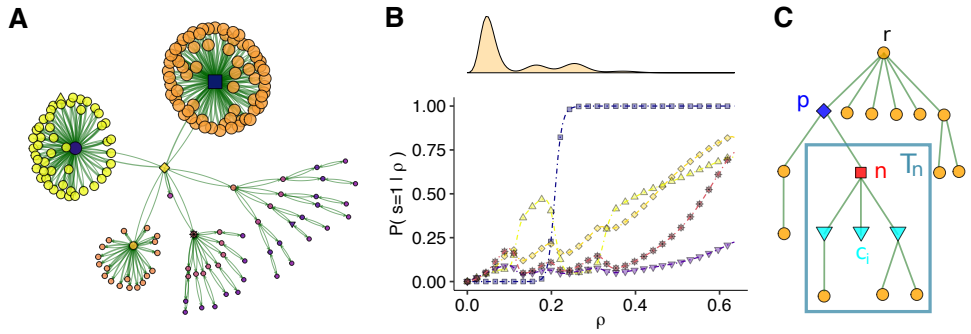

Figure 3: A: Artificial tree for SD. Link strengths are proportional to the ICM weights, node sizes to the activation probabilities as computed by BP. B: Conditional activation probability of nodes matching the symbols in A corresponding to the cascade size distribution on top. Fifteen distinct clusters are represented by different node colors. C. Illustration of variables in SDP.

The Independent Cascade Model

The ICM models the binary, stochastic, and discrete activation dynamics of nodes in an undirected network $G = (V, E)$ consisting of $N = |V|$ nodes that are connected by links in E. Each node i is either inactive ($s_i = 0$) or active ($s_i = 1$) and can only switch from an inactive to an active state, but not vice versa. Initially, each node i activates with probability p_i independently of the other nodes. In the next time step ($t = 1$), an active node i can trigger new activations of its neighbors $j \in \mathrm{nb(i)} := \{\mathrm{j} \mid (\mathrm{i,j}) \in \mathrm{E}\}$. Its degree $d_i = |\mathrm{nb(i)}|$ counts the number of neighbors it has. Each neighbor j activates independently with probability w_{ij} and can cause new activations in the next time step. This way, a cascade keeps propagating, where several activations can happen at the same time t and each node becoming active at t can trigger new activations only in the next time step $t + 1$ but not any later times. The process ends at time $T \leq N$, when no further activations can occur. Then, the fraction of active nodes $\rho = 1/N \sum_{i=1}^{N} s_i(T)$ defines the final cascade size. This is the realization of a random variable C with probability distribution $p_C(\rho)$, as the cascade process is stochastic. $p_C(\rho)$ is also termed probability mass function of C and has support $\{0, 1/N, ..., 1\}$. In summary, an ICM is parametrized by $(\mathbf{p}, \mathbf{W})$, where the vector $\mathbf{p}$ has components p_i and the matrix $\mathbf{W}$ entries w_{ij}. Note that W can also refer to network weights and encode directedness by $w_{ij} \neq w_{ji}$ and $w_{ij} = 0$. If the vector $\mathbf{p}$ has only binary components $p_i \in \{0, 1\}$, we call nodes with entry 1 the seeds $S = \{i \mid p_i = 1\}$, as they are activated initially. We denote the corresponding cascade size with ρ_S.

Algorithmic approach

Fig. 2 in the supplementary material gives an overview of our main contribution: four variants of a message passing algorithm. The core is formed by *Subtree Distribution Propagation (SDP)*, which computes the final cascade size distribution based on a rooted tree and ICM as input. It is efficient, as it needs to visit each node only once, and can be parallelized according to the tree structure. Starting in the leaves (that is, nodes with degree $d_n = 1$ at the "periphery" of the network), each node n sends information about the cascade size distribution of the subtree T_n rooted in n to its parent p. It only requires information by its children as input. Finally, the output is constructed in the root. The relevant variables are visualized in Fig. 3 C. To compute the activation probability of nodes conditional on the final cascade size, *Conditional Subtree Distribution Propagation (conSDP)* adds a backpropagation step to SDP. Information is sent from the root to its children until it reaches the leaves. Thus, each node is visited twice in total. SDP and conSDP are exact, but are limited to trees. To obtain an approximate variant that applies to any (simple) network G, we introduce *(conditional) Tree Distribution Approximation (TDA)* as extension of (conditional) SDP. As in (Burkholz 2019), the goal is to reduce G to a tree M (for instance a maximum spanning tree (MST)) and to run SDP (or conSDP) on M. However, to compensate for the removal of edges and thus dependencies of node activations, we increase the initial activation probabilities $\mathbf{p}$ of the ICM. There, we assume that former neighbors j have activated independently before a node i with a probability p_{ji} that we estimate by Belief Propagation (BP). When BP does not converge, we could substitute another approach such as the junction tree algorithm. For simplicity and computational efficiency, we will only consider BP. In the following, we detail the information propagation equations of the respective algorithms.

Subtree Distribution Propagation

Theorem 1 (SDP)**.** *The final cascade size distribution $p_C(\rho)$ of an ICM $(\mathbf{p}, \mathbf{W})$ on a tree G with root r and N nodes is given as output of the following message passing algorithm.*
Starting in the leaves, each node n sends the functions $p_{B_n}(t)$, $p_{A_n^0}(t)$, and $p_{A_n^\Sigma(t)}$ for $t = 0, ..., N$ as messages to its parent p. We have

$$p_{B_n}(0) = 1 - p_n, \qquad p_{B_n}(1) = p_n(1 - w_{np}),$$
$$p_{A_n^0}(0) = (1 - p_n)(1 - w_{pn}), \qquad p_{A_n^0}(1) = (1 - p_n)w_{pn} + p_n(1 - w_{np}),$$
$$p_{A_n^\Sigma}(0) = (1 - p_n)(1 - w_{pn}), \qquad p_{A_n^\Sigma}(1) = (1 - p_n)w_{pn} + p_n.$$

for a leave n (with degree $d_n = 1$). Otherwise, define for a

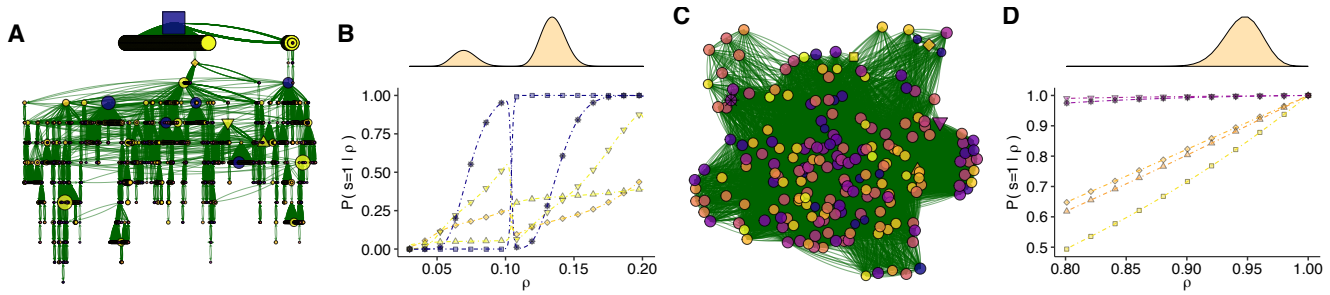

Figure 4: A: Corporate ownership network for SD. Link strengths are proportional to the ICM weights, node sizes to the activation probabilities as computed by BP. B: Conditional activation probability of exemplary nodes matching the symbols in A corresponding to the cascade size distribution on top. 40 distinct clusters are represented by different node colors. C. miRNA network for gastrointestinal cancer with 20 clusters. D. Conditional activation probabilities for selected miRNAs.

node n with k children $c_1, ..., c_k$:

$$p_{nf}(t) = (1-p_n)p_{B_{c_1}} * \cdots * p_{B_{c_k}}[t],$$
$$p_{lf}(t) = (1-p_n)p_{A^0_{c_1}} * \cdots * p_{A^0_{c_k}}[t],$$
$$p_f(t) = p_{A^\Sigma_{c_1}} * \cdots * p_{A^\Sigma_{c_k}}[t] - p_{lf}(t),$$

where $*$ denotes a convolution. Then, an intermediate node $n \neq r$ ($d_n > 1$) with $k_n = d_n - 1$ children, sends the messages:

$$\begin{aligned} p_{B_n}(t) =& p_{nf}(t) + (1-w_{np})p_f(t-1), \\ p_{A^0_n}(t) =& (1-w_{pn})p_{nf}(t) + w_{pn}p_{lf}(t-1) \\ &+ (1-w_{np})p_f(t-1), \\ p_{A^\Sigma_n}(t) =& p_{A^0_n}(t) + w_{np}p_f(t-1). \end{aligned}$$

In the root, $n = r$ with $k_r = d_r$ children, the final cascade size distribution is calculated as: $p_C(t/N) = p_{nf}(t) + p_f(t-1)$.

Proof Outline. The derivation of SDP relies solely on combinatorial arguments, where the probabilities of all possible events are added, respecting the right order of activations. Each node n receives messages by its children, combines those to new messages that provide information about the number of activated nodes t in the subtree T_n rooted in n with assumptions on the parent's state s_p, and sends these messages to its parent p. Fig. 3 C visualizes the relevant variables. p_{B_n} accounts for the cases when n does not trigger the activation of its parent, while $p_{A^\Sigma_n}$ sums up the cases where the parent p activates at the same time as n or before. $p_{A^0_n}$ is auxiliary to subtract an infeasible case when the messages received by children c_i are combined and added. This combination is facilitated by convolutions, as the subtree cascade size distributions rooted in the children are independent for a fixed parent state s_p. As s_n is binary, we distinguish two main cases for the subtree cascade size distributions rooted in n. A third one is auxiliary to respect the right order of activations. p_{nf} refers to "no activation" of n, p_{lf} to "late activation" (or "no initial activation"), and p_f to all cases when n activates. p_{lf} is auxiliary to subtract from p_f the case that no child actually triggers the activation of n and n does not activate initially even though the children assume an activated parent n. Finally, the cascade size distribution is computed in the root (where no parent state needs to be considered). In case of t final activations, either the root does not activate with $p_{nf}(t)$ or activates with $p_f(t-1)$ so that $t-1$ other nodes activate. □

Note that all operations can be performed in Fourier space. Thus, the convolutions simplify to elementwise multiplication of vectors. Note that only at the root, in the end, we have to compute a single inverse Fourier transformation to have a probability distribution. The exact algorithm in Fourier space and a more detailed proof are provided in the supplementary material.

Tree Distribution Approximation To apply the same principle to a general network $G = (V, E)$ and thus SDP to a maximum spanning tree $M = (V_M, E_M)$ of G, we have to regard the direct influence of neighbors $\mathrm{dn(i)} = \{\mathrm{j} \in \mathrm{V} \mid (\mathrm{i,j}) \in \mathrm{E},\ (\mathrm{i,j}) \notin \mathrm{E_M}\}$ a node i is not connected with any more in M. For all nodes, those can be estimated by BP.

Belief Propagation for ICM Given an ICM $(\mathbf{p}, \mathbf{W})$ on a network G, the probability $p_{ij} = \mathbb{P}(s_i = 1 \parallel s_j = 0)$ that i activates without j's contribution is estimated iteratively by repeating R times

$$Q_i = (1-p_i) \prod_{n \in \mathrm{nb(i)}} (1 - w_{ni}p_{ni}), \quad i = 1, ..., N,$$
$$p_{ij} = 1 - \frac{Q_i}{1 - w_{ji}p_{ji}}, \quad i, j = 1, ..., N,$$

where we initialize with the initial ICM activation probability $p_{ij} = p_i$.

p_{ij} for $j \in \mathrm{dn(i)}$ are used next to adapt the initial activation probabilities $\mathbf{p}^{(\mathrm{M})}$ of an ICM on M (instead of G). They incorporate the influence of deleted neighbors by assuming that they activate independently before i with p_{ji}.

ICM on MST The ICM $(\mathbf{p}^{(\mathbf{M})}, \mathbf{W}^{(\mathbf{M})})$ on M is given as

$$p_i^{(M)} = 1 - (1 - p_i) \prod_{j \in \mathrm{dn(i)}} (1 - p_{ji}), \;\; i = 1, ..., N,$$

$$w_{ij}^{(M)} = w_{ij} \text{ for } (\mathrm{i}, \mathrm{j}) \in \mathrm{E_M}, \text{ and } \mathrm{w_{ij}^{(M)}} = 0 \text{ otherwise.}$$

Conditional activation probability

The activation probability of a node conditional on the final cascade size is straight forward to compute for the root at the end of SDP (see Thm. 1) as $\mathbb{P}(s_r = 1 \mid C = t/N) = p_f(t-1)/p_C(t/N)$. Yet, to obtain the same for every other node n, we have to calculate additional information to treat them as a root. After SDP, only messages from the former parent are missing, so that $p_{B_p^c}(t)$, $p_{A_p^{0c}}(t)$, and $p_{A_p^{\Sigma c}(t)}$, where p is treated as a child, while n is the new parent. Thus, starting in the root r, each parent p backpropagates messages to their children n, where $\mathbb{P}(s_n = 1 \mid C = t/N)$ can be computed.

BackPropagation for ConSDP (or ConTDA) Using the notation of Thm. 1, each parent p sends to its child n:

$$\begin{aligned}
p_{B_p^c}(t) =& p_{nf\setminus n}^{(p)}[t] + (1 - w_{pn}) p_{f\setminus n}^{(p)}(t-1), \\
p_{A_p^{0c}}(t) =& (1 - w_{np}) p_{nf\setminus n}^{(p)}(t) + w_{np} p_{lf\setminus n}^{(p)}(t-1) \\
&+ (1 - w_{pn}) p_{f\setminus n}^{(p)}(t-1), \\
p_{A_p^{\Sigma c}}(t) =& p_{A_p^{0c}}(t) + w_{np} p_{f\setminus n}^{(p)}(t-1),
\end{aligned}$$

where the contribution of n in the convolution forming $p_x^{(p)}$ is removed (for $x \in \{nf, lf, f\}$). Note the swap of n and p in comparison with Thm. 1. For all neighbors, children and parent, the messages from SDP are combined with the new one received by the parent as:

$$\begin{aligned}
&p_{nf}^{(n)}(t) = p_{nf} * p_{B_p^c}[t], \;\; p_{lf}^{(n)}(t) = p_{lf} * p_{A_p^{0c}}[t], \\
&p_f^{(n)}(t) = (p_f + p_{lf}) * p_{A_p^{\Sigma c}}[t] - p_{lf}^{(n)}(t)
\end{aligned}$$

and form the conditional activation probability of n as

$$\mathbb{P}(s_n = 1 \mid C = t/N) = \left(1 + \frac{p_{nf}^{(n)}(t)}{p_f^{(n)}(t-1)}\right)^{-1}.$$

The precise algorithm is stated in the supplementary material.

Algorithmic complexity

The core algorithm, SDP (and backpropagation), is quite efficient, as each node is visited only once (or twice with backpropagation). The total number of performed operations (without parallelization) is $O(N^2)$. Yet, an approximate version that computes $p_C(\rho)$ for a finite resolution (for instance on an equidistant grid of $[0, 1]$) would reduce the complexity to $O(N)$. With parallelization of computations in nodes that have received the messages by their children, this can be even brought down to $O(h)$, meaning that it is linear in the height of the input tree. The bottleneck of TDA is thus given by BP to preprocess the ICM before SDP can be employed. Each node has to be visited several times. Neverthless, BP can easily be parallelized within a message passing framework, greatly increasing its potential utility.

Influence maximization

We illustrate the utility of these algorithms by influence maximization, a common optimization scenario that usually requires sampling. The goal is to maximize an influence function $\sigma(S)$ by choice of initial seed set S subject to a constraint on its cardinality $|S| = k$. For the original choice $\sigma(S) = \mathbb{E}(\rho_S)$, this problem is NP-hard (Kempe, Kleinberg, and Tardos 2003), but can be well approximated, since $\sigma(S)$ is submodular. Greedy seed size selection is guaranteed to achieve a solution of at least $1 - 1/e - \epsilon$ of the objective reached by the optimum. Yet, if we are interested in the right tail of the cascade size distribution and adjust $\sigma(S)$ accordingly, we usually loose submodularity. We discuss two response function choices. A common tail measure in systemic risk analysis is Expected Shortfall, also known as conditional Value at Risk $\mathbb{E}(\rho_S \mid \rho_S \geq \mathrm{VaR}_\alpha(\rho_\mathrm{S}))$, where the Value at Risk is defined as $\mathrm{VaR}_\alpha(\rho_\mathrm{S}) = \max\{\mathrm{x} \mid \mathbb{P}(\rho_\mathrm{S} \geq \mathrm{x}) \geq \alpha\}$. It measures the average cascade size for the best $alpha$ cases and is not submodular (Maehara 2015). Alternatively, we can give more weight to the tail when taking an average. For any non-negative, measurable function $f : \{0, 1/N, ..., 1\} \to \mathbb{R}_+$ the response function $\sigma(S) = \mathbb{E}(f(\rho_S)) = \sum_{r=0}^{N} p_C(\rho_S = r/N) f(r/N)$ is positive, monotonously increasing, but in general not submodular. Consequently, we do not have the same guaranties for the Greedy algorithm as in the original problem. In our experiments, however, a Greedy search still shows good performance when compared with exhaustive search.

Numerical Experiments

We perform experiments on three different networks: an artificial tree consisting of $N = 181$ nodes (Fig. 3), a locally tree-like real world network of corporate ownership relationships (Norlen et al. 2002) with $N = 4475$ (Fig. 4 A), and a dense correlation network of miRNA expression profiles using data from gastrointestinal cancer (Nalluri et al. 2017) with $N = 201$ nodes (Fig. 4 C). The first two are unweighted, while the miRNA network has estimated ICM weights w_{ij}. For the first two, we can choose the weights according to different cascade models.

We discuss three spreading mechanisms to demonstrate the high variability in possible outcomes. The exposure diversification (ED) model follows similar patterns as the well studied threshold model (Granovetter 1978; Watts 2002) with $w_{ij} = 0.05 + 0.5/d_j$. High degree nodes are more difficult to activate by single network neighbors. In contrast, the damage diversification (DD) model (Burkholz, Garas, and Schweitzer 2016) makes high degree nodes less likely to infect a network neighbor. We study a simplified version with $w_{ij} = 0.6$ for $d_i \geq d_j$ and $w_{ij} = 0.8$ otherwise, as this shows interesting patterns for influence maximization. Unless stated otherwise, the initial activation probability of each node is $p_i = 0.05$ and for miRNAs $p_i = 0.01$ to mitigate extensive spreading due to the high network density. Last, we study a social dynamics (SD) model with $w_{ij} = 0.05 + 0.5 d_i/d_j/Z$ and $Z = \max_{i,j}(d_i/d_j)$. This choice reflects the intuition that well regarded nodes, like big news sites, have a high degree and are more influential.

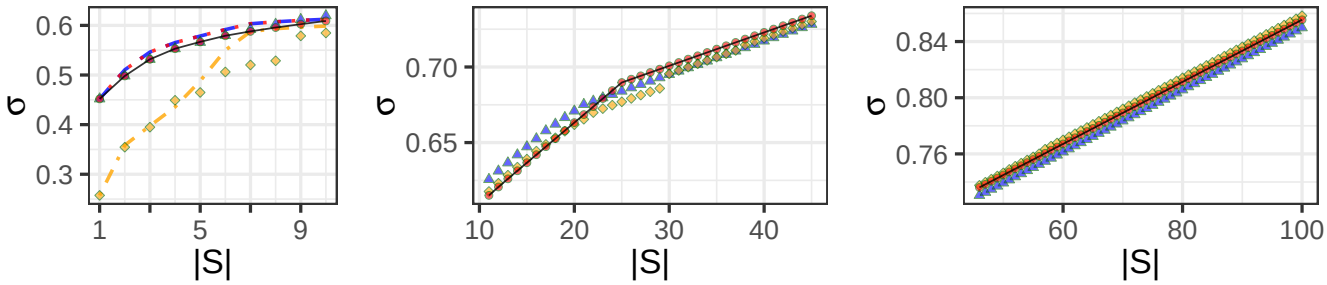

Figure 5: Average cascade size $\sigma(S) = \mathbb{E}(\rho_S)$. $|S|$ seeds are chosen according to different Greedy objectives: average cascade size (red circles and black line), exponential weights (blue triangles), expected shortfall (yellow squares). Lines in different colors belong to exhaustive maximization of the corresponding objective based on node clusters.

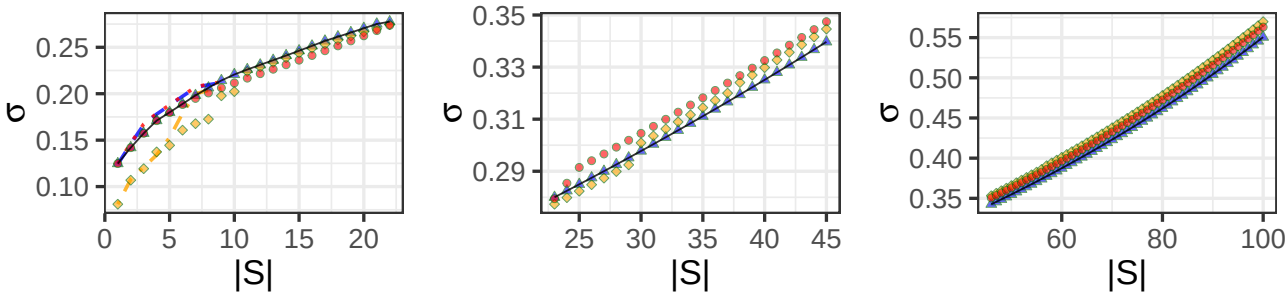

Figure 6: Exponentially weighted average cascade size $\sigma(S) = \mathbb{E}(f(\rho_S))$. $|S|$ seeds are chosen according to different Greedy objectives: average cascade size (red circles), exponential weights (blue triangles and black line), expected shortfall (yellow squares). Lines in different colors belong to exhaustive maximization of the corresponding objective based on node clusters.

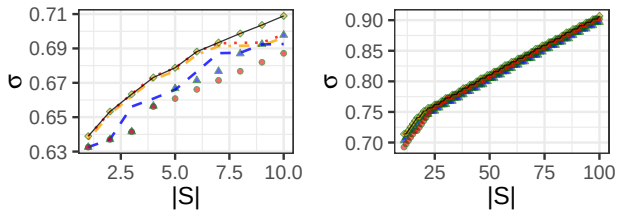

Figure 7: Expected shortfall. $|S|$ seeds are chosen according to different Greedy objectives: average cascade size (red circles), exponential weights (blue triangles), expected shortfall (yellow squares and black line). Lines in different colors belong to exhaustive maximization of the corresponding objective based on node clusters. $\sigma(S) = \mathbb{E}(\rho_S \mid \rho_S \geq \mathrm{VaR}_{0.05}(\rho_S))$.

However, they can be difficult to convince to spread information.

First, we demonstrate that SDP and TDA work correctly by comparison with the empirical distribution obtained by sampling (through Monte Carlo simulation) from the true model 10^6 times. Belief propagation iterations are always repeated $R = 10$ times. Fig. 2 shows perfect agreement for the tree and locally tree-like ownership network. Even for the dense miRNA network, TDA performs surprisingly well, but we cannot expect this good agreement in general. The mean squared error for the tree is always smaller than 10^{-8}, ca. $5{\cdot}10^{-8}$ for the corporate ownership, and $4.3{\cdot}10^{-6}$ for the miRNA network. In general, we note that for realistic model parameter choices, we find broad and multi-modal cascade size distributions. These underline the importance to regard full cascade size distributions rather than averages.

Conditional activation probabilities

The activation probabilities of exemplary nodes are shown in Figs. 3B, 4B, 4D as obtained by ConSDP or ConTDA. They vary substantially in dependence on the cascade size and often increase non-monotonically, different from what one might expect. Big hubs are more predictable and are always active above a certain cascade size. Their activation usually marks larger cascades but does not explain the largest. These only occur with the activation of nodes that are more difficult to reach by cascades.

Nodes, which are topological interchangeable, also have an identical conditional activation probability. The identification of such symmetries is particularly interesting in the analysis of biological networks like the cancer related miRNA network, since these hint towards similar functions of nodes within pathways and thus redundancies in the network. In addition, similar conditional activation probabili-

ties translate into similar effects as seeds. This is relevant information if we want to optimize such effects. Therefore, we cluster nodes i based on vectors v^i with components $v^i_{r+1} = \mathbb{P}(s_i = 1 \mid \rho = r/N)$. Any clustering algorithm could be employed. For simplicity, we choose kmeans. Node colors in Fig. 3-4 indicate the cluster membership of a node.

Influence maximization

Assigning nodes to clusters reduces the space of seeds we need to consider in influence maximization considerably. Furthermore, we can use SDP and TDA to compute the final cascade size efficiently and accurately without the need to rely on sampling. In the main paper, we focus on the DD ICM and the tree. Corresponding results for the other networks are provided in the supplementary material. We have three different objectives to compare, the average cascade size (Fig. 5), a weighted average of the cascade size with $f(\rho) = \exp(4\rho)/\exp(4)$ (Fig. 6) that assigns a higher weight to the tail, and expected shortfall with $\alpha = 0.05$ (Fig. 7). Each objective is optimized in a Greedy approach. Hence, in each step, we add a node n to the current seed set S so that $n = \mathrm{argmax}_{\mathrm{v}\in\mathcal{C}}\sigma(\mathrm{S}\cup\{\mathrm{v}\})$, where we only need to compare different cluster representatives v in $\mathcal{C}$. For each objective, we find different Greedy optimizing seed sets. Interestingly, some of these sets achieve a higher influence than the seeds optimized for this case, as also noted in the context of portfolio optimization (Ohsaka and Yoshida 2017). For instance, for a high number of seeds ($45 < |S| < 100$), seeds based on expected shortfall perform best for all three objectives. For a small number of initial seeds, however, seeds for expected shortfall perform poorly for other objectives than expected shortfall. Yet, they consistently provide the best results for expected shortfall itself. For a small number of seeds, seeds optimized for the weighted average can outperform the original average objective. Yet, the Greedy results in general could be far off the true optima. To test this, we perform an exhaustive search based on cluster representatives for up to seed size 10. We only allow for up to three members of the same one clusters. Therefore, we do not regard sets consisting of several seeds belonging to the same cluster for several clusters. These are unlikely optimizing sets for small seed set sizes in an event. In this way, we verify that the performance difference between a Greedy and exhaustive search is not substantial, at least for a small number of seeds. In addition, clustering of nodes enables an exhaustive search for a far higher number of seeds than without clustering.

Discussion & Outlook

The core algorithms that we have derived, SDP and conSDP, compute the final cascade size distribution given a general independent cascade model and can therefore replace expensive sampling procedures in optimization problems such as influence maximization. These algorithms are exact on trees. Real world networks usually have additional connections that can create multiple short loops. These introduce stronger dependencies of activations so that with higher probability nodes activate either together or not at all. By approximating a denser network with a tree in TDA, we treat some nodes as conditionally independent when they are not and thus underestimate the probability that they activate (or do not activate) together. As a consequence, we potentially underestimate the variance of the cascade size distribution. Furthermore, we can interpret TDA as variational approach to obtain a proxy for the final cascade size distribution. BP can also be substituted by a more reliable approach (for instance the junction tree algorithm) in case higher accuracy is needed.

Actually, the SDP principle generalizes much further. It only requires estimates of cascade size distributions for subgraphs (considering states of parents). These can also be obtained by Monte Carlo simulations or a combination of sampling and SDP. As long as such subgraphs are connected like trees, SDP can be used to speed up sampling (and reduce the number of necessary realizations), and simulations on subnetworks can be used to improve the accuracy of TDA.

Cascade size distributions are not only important to assess the robustness of a network and the risk of extreme events. Conditional activation probabilities are also indicative of the role that specific nodes play in spreading a cascade. This provides information for policy makers or optimization algorithms to influence the outcome of a cascade. As a demonstration of the method, we used a simple Greedy strategy, but more sophisticated IM-like algorithms that rely on sampling would also benefit from TDA. Furthermore, TDA can be easily generalized to cover variations of the IM problem, where nodes are weighted by positive $\beta_i \geq 0$ and the distribution of $\sum_i \beta_i s_i(T)$ is of interest. Our algorithms can also be extended to allow for distributions on the ICM parameters and could therefore aid robust influence maximization under model parameter uncertainty (Kalimeris, Kaplun, and Singer 2019).

Conclusion

We have derived efficient message passing algorithms to compute the final cascade size distribution $p_C(\rho)$ for a given finite network and Independent Cascade Model. The core is provided by Subtree Distribution Propagation, which is exact on trees and requires $O(N^2)$ operations, where N denotes the number of network nodes.

For an arbitrary network, we have introduced Tree Distribution Approximation (TDA), which relies on Belief Propagation as preprocessing and approximates the final cascade size distribution. It is accurate on locally tree-like networks and shows surprisingly good performance on an exemplary dense network of miRNA associated with gastrointestinal cancer.

In addition, we can compute the activation probabilities of nodes conditional on the final cascade size. These are particularly informative in systemic risk analyses, as they allow the focus on extreme events. We have used these probabilities to cluster nodes with similar functionality. The resulting group representatives reduce the number of seed candidates in Greedy influence maximization. We have further demonstrated the usefulness of our algorithms to optimize alternative influence functions like expected shortfall.

Acknowledgments

RB and JQ were supported by a grant from the US National Cancer Institute (1R35CA220523).

References

[Abebe, Adamic, and Kleinberg 2017] Abebe, R.; Adamic, L. A.; and Kleinberg, J. M. 2017. Mitigating overexposure in viral marketing. *CoRR* abs/1709.04123.

[Battiston et al. 2016] Battiston, S.; Caldarelli, G.; May, R. M.; Roukny, T.; and Stiglitz, J. E. 2016. The price of complexity in financial networks. *Proceedings of the National Academy of Sciences* 113(36):10031–10036.

[Budak, Agrawal, and El Abbadi 2011] Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, 665–674. New York, NY, USA: ACM.

[Burkholz, Garas, and Schweitzer 2016] Burkholz, R.; Garas, A.; and Schweitzer, F. 2016. How damage diversification can reduce systemic risk. *Physical Review E* 93:042313.

[Burkholz, Herrmann, and Schweitzer 2018] Burkholz, R.; Herrmann, H. J.; and Schweitzer, F. 2018. Explicit size distributions of failure cascades redefine systemic risk on finite networks. *Scientific Reports* 1–8.

[Burkholz 2019] Burkholz, R. 2019. Efficient message passing for cascade size distributions. *Scientific Reports*.

[Domingos and Richardson 2001] Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, 57–66. New York, NY, USA: ACM.

[Du et al. 2014] Du, N.; Liang, Y.; Balcan, M.; and Song, L. 2014. Influence function learning in information diffusion networks. In Xing, E. P., and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 2016–2024. Bejing, China: PMLR.

[Farajtabar et al. 2015] Farajtabar, M.; Gomez-Rodriguez, M.; Zamani, M.; Du, N.; Zha, H.; and Song, L. 2015. Back to the past: Source identification in diffusion networks from partially observed cascades. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*.

[Garimella et al. 2017] Garimella, K.; Gionis, A.; Parotsidis, N.; and Tatti, N. 2017. Balancing information exposure in social networks. In *NIPS*, 4666–4674.

[Gleeson and Porter 2018] Gleeson, J. P., and Porter, M. A. 2018. Complex spreading phenomena in social systems. In Jørgensen, S., and Ahn, Y.-Y., eds., *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*. Cham: Springer. 81–95.

[Goldenberg, Libai, and Muller 2001] Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12(3):211–223.

[Granovetter 1978] Granovetter, M. S. 1978. Threshold Models of Collective Behavior. *The American Journal of Sociology* 83(6):1420–1443.

[Kalimeris, Kaplun, and Singer 2019] Kalimeris, D.; Kaplun, G.; and Singer, Y. 2019. Robust influence maximization for hyper-parametric models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 3192–3200.

[Kempe, Kleinberg, and Tardos 2003] Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 137–146. New York, NY, USA: ACM.

[Kempe, Kleinberg, and Tardos 2005] Kempe, D.; Kleinberg, J.; and Tardos, E. 2005. Influential nodes in a diffusion model for social networks. In *Proceedings of the 32Nd International Conference on Automata, Languages and Programming*, ICALP'05, 1127–1138. Berlin, Heidelberg: Springer-Verlag.

[Leskovec, Adamic, and Huberman 2007] Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Trans. Web* 1(1).

[Leskovec et al. 2007] Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, 420–429. New York, NY, USA: ACM.

[Maehara 2015] Maehara, T. 2015. Risk averse submodular utility maximization. *Oper. Res. Lett.* 43(5):526–529.

[McNeil, Frey, and Embrechts 2015] McNeil, A. J.; Frey, R.; and Embrechts, P. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools Revised edition*, volume 1. Princeton University Press, 2 edition.

[Morris 2000] Morris, S. 2000. Contagion. *The Review of Economic Studies* 67(1):57–78.

[Nalluri et al. 2017] Nalluri, J.; Rana, P.; Barh, D.; Azevedo, V.; Dinh, T.; Vladimirov, V.; and Ghosh, P. 2017. Determining causal mirnas and their signaling cascade in diseases using an influence diffusion model. *Scientific Reports* 7.

[Newman 2002] Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Physical Review E* 66:016128.

[Norlen et al. 2002] Norlen, K.; Lucas, G.; Gebbie, M.; and Chuang, J. 2002. EVA: Extraction, Visualization and Analysis of the Telecommunications and Media Ownership Network.

[Ohsaka and Yoshida 2017] Ohsaka, N., and Yoshida, Y. 2017. Portfolio optimization for influence spread. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 977–985. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

[Tschiatschek et al. 2018] Tschiatschek, S.; Singla, A.; Gomez Rodriguez, M.; Merchant, A.; and Krause, A. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, 517–524. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

[Vosoughi, Roy, and Aral 2018] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

[Watts 2002] Watts, D. J. 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences USA* 99:5766–5771.

[Wen et al. 2017] Wen, Z.; Kveton, B.; Valko, M.; and Vaswani, S. 2017. Online influence maximization under independent cascade model with semi-bandit feedback. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 3022–3032.

[Xu and Chen 2015] Xu, W., and Chen, H. 2015. Scalable rumor source detection under independent cascade model in online social networks. In *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, 236–242.

[Y. Lokhov and Saad 2016] Y. Lokhov, A., and Saad, D. 2016. Optimal deployment of resources for maximizing impact in spreading processes. *Proceedings of the National Academy of Sciences* 114.

[Zhu, Chen, and Ying 2017] Zhu, K.; Chen, Z.; and Ying, L. 2017. Catch'em all: Locating multiple diffusion sources in networks with partial observations. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 1676–1683.