Action selection in growing state spaces: Control of Network Structure Growth

Dominik Thalmeier[1], Vicenç Gómez [2], and Hilbert J. Kappen [1]

[1]Donders Institute for Brain, Cognition and Behaviour
Radboud University Nijmegen, the Netherlands
[2]Department of Information and Communication Technologies
Universitat Pompeu Fabra. Barcelona, Spain

Abstract

The dynamical processes taking place on a network depend on its topology. Influencing the growth process of a network therefore has important implications on such dynamical processes. We formulate the problem of influencing the growth of a network as a stochastic optimal control problem in which a structural cost function penalizes undesired topologies. We approximate this control problem with a restricted class of control problems that can be solved using probabilistic inference methods. To deal with the increasing problem dimensionality, we introduce an adaptive importance sampling method for approximating the optimal control. We illustrate this methodology in the context of formation of information cascades, considering the task of influencing the structure of a growing conversation thread, as in Internet forums. Using a realistic model of growing trees, we show that our approach can yield conversation threads with better structural properties than the ones observed without control.

Keywords: control, complex Networks, sampling, conversation threads

1 Introduction

Many complex systems can be described as dynamic processes which are characterized by the topology of an underlying network. Examples of such systems are human interaction networks, where the links may represent transmitting opinions Olfati-Saber et al. [2007], Dai and Mesbahi [2011], Centola and Baronchelli [2015], habits Centola [2010], Farajtabar et al. [2014], money Gai and Kapadia [2010], Amini et al. [2016], Giudici and Spelta [2016] or viruses Pastor-Satorras and Vespignani [2001], Eguíluz and Klemm [2002]. Being able to control, or just influence in some way, the dynamics of such complex networks may lead to important progress, for example, avoiding financial crises, preventing epidemic outbreaks or maximizing information spread in marketing campaigns.

The control of the dynamics on networks is a very challenging problem that has attracted significant interest recently Liu et al. [2011], Cornelius et al. [2013], Gao et al. [2014], Yan et al. [2015]. Existing approaches typically consider network controllability as the controllability of

the dynamical system induced by the underlying network structure. While it is agreed that network controllability critically depends on the network structure, the problem of how to control the network structure itself while it is evolving remains open.

The network structure is determined by the dynamics of addition and deletion of nodes and links over time. In this paper, we address the problem of influencing this dynamics in the framework of stochastic optimal control. The standard way to address these problems is through the Bellman equation and dynamic programming. Dynamic programming is only feasible in small problems and requires approximations when the state and action spaces are large. In the setting of network growth, this problem is more severe, since the state space increases (super-)exponentially with the number of nodes.

In order to deal with this curse of dimensionality, we propose to approximate the network growth control problem by a special class of stochastic optimal control problems, known as Kullback-Leibler (KL) control or Linearly-Solvable Markov Decision Problems (LMDPs) Todorov [2009], Kappen et al. [2012]. For this class of problems, one can use efficient adaptive importance sampling methods that scale well in high dimensions. The optimal solution for the KL-control problem tends to be sparse, so that only a few next states become relevant, effectively reducing the branching factor of the original problem. The obtained solution of the KL-control problem is then used to compute the optimal action in the original problem that does not belong to the KL-control class.

In the next section we present our proposed general methodology. We then apply it to a realistic problem: influencing the growth process of cascades in online forums, in order to maximize structural network measures that are connected to the quality of an online conversation thread. We conclude the paper with a discussion.

2 Optimal Network Growth as a Control Problem

We now formulate the network growth control problem as a stochastic optimal control problem. Let $x_t \in \mathcal{X}$, with $\mathcal{X}$ being the set of all possible network structures, denote the growing structure (state) of the network at time t and let $P(x'|x,u)$ describe the network dynamics, where the control variable $u \in \mathcal{U}$ denotes possible actions we can perform in order to manipulate the network. Let us label the default action, which means not interacting with the system, with $u = 0$. We denote the corresponding dynamics without control as the uncontrolled process $p(x'|x) := P(x'|x, u = 0)$.

At each time-step t, we incur an arbitrary cost function on the network state $r(x,t)$ which is assigned when the state is reached. The state cost $r(x,t)$ penalizes network structures that are not convenient in the particular context under consideration. For example, if one wants to favour networks with large average clustering coefficient $C(x)$, then $r(x,t) = -C(x)$. Alternatively, one can consider more complex functions, such as the structural virality or Wiener index Mohar and Pisanski [1988], as proposed recently Goel et al. [2015], to maximize the influence in a social network. In general, any measure that can be (efficiently) computed from x fits the presented framework.

Our objective is to find the control function $u(x,t) : \mathcal{X} \times \mathbb{N} \mapsto \mathcal{U}$ which minimizes the total cost over a time horizon T, starting at state x at initial time $t = 0$

$$\mathcal{C}\left(x, t = 0, u(\cdot)\right) = r\left(x, 0\right) + \left\langle \sum_{t'=t+1}^{T} r\left(x_{t'}, t'\right) \right\rangle_{P(x_{1:T}|x,u(\cdot),t=0)}, \tag{1}$$

where the expectation is taken with respect to the probability $P(x_{1:T}|x, u(\cdot), t = 0)$ over paths $x_{1:T}$ in the state space, given state x at time $t = 0$ using the control-function $u(\cdot)$. The probability of a path is given by $P(x_{t+1:T}|x, u(\cdot), t = 0) = \prod_{s=t}^{T-1} P(x_{s+1}|x_s, u(x_s, s), s)$.

Computing the optimal control can be done by dynamic programming Bertsekas [1995]. We introduce the optimal cost-to-go

$$J(x, t) = \min_{u(\cdot)} \mathcal{C}\left(x, t, u(\cdot)\right), \tag{2}$$

which is an expectation of the cumulative cost starting at state x and time t and acting optimally thereafter. This can be computed using the Bellman equation

$$J(x, t) = \min_{u} \left(r(x, t) + \left\langle J(x', t + 1) \right\rangle_{P(x'|x,u,t)} \right). \tag{3}$$

From $J(x, t)$, the optimal control is obtained by a greedy local optimization:

$$u^*(x, t) = \text{argmin}_u \left(r(x, t) + \left\langle J(x', t + 1) \right\rangle_{P(x'|x,u,t)} \right). \tag{4}$$

In general, the solution to equation (3) can be computed recursively using dynamic programming Bertsekas [1995] for all possible states. This is however infeasible for controlling network growth, as the computation is of polynomial order in the number of states and the state space of networks increases super-exponentially on the number of nodes. E.g. for directed unweighed networks, there are 2^{N^2} possible networks with N labelled nodes.

3 Approximating the network growth problem by a Kullback-Leibler control problem

In this section we present our main approach, which first computes the optimal cost-to-go on a relaxed problem and then uses it as a proxy for the original optimal cost-to-go. In the next subsection, we introduce the class of KL-control problems that we use as a relaxation. We then illustrate KL-control using a tractable example of tree growth. In subsection 3.3, we explain how can we approximate the KL-control solution using the cross-entropy method. Finally, in subsection 3.4 we show how can we use that result to compute the action selection in the original problem.

3.1 Kullback-Leibler control

In order to efficiently compute the optimal cost-to-go, we make the assumption that our controls directly specify the transition probabilities between two subsequent network structures, e.g. $P(x'|x, u(t)) \approx u(x'|x, t)$. Further, we define the natural growth process of the network (the uncontrolled dynamics) as a Markov chain with transition probabilities $p(x'|x)$. Because our influence on the network dynamics is limited, we add a regularization term to the total cost defined in equation (1) that penalizes deviations from $p(x'|x)$. The approximated control cost becomes

$$\mathcal{C}_{\text{KL}}^{\lambda}\left(x, t, u(\cdot)\right) = \lambda \text{KL}\left[u\left(x_{t+1:T}|x, t\right) \parallel p\left(x_{t+1:T}|x, t\right)\right] + r\left(x, t\right) + \left\langle \sum_{t'=t+1}^{T} r\left(x_{t'}, t'\right) \right\rangle_{u(x_{t+1:T}|x,t)}, \tag{5}$$

with the KL-divergence

$$\mathrm{KL}\left[\mathrm{u}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right) \parallel \mathrm{p}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)\right] = \left\langle \log \frac{\mathrm{u}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)}{\mathrm{p}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)} \right\rangle_{\mathrm{u}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t})},$$

which measures the closeness of the two path distributions, $\mathrm{p}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)$ and $\mathrm{u}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)$. The parameter λ thereby regulates the strength of this penalization.

With this assumption, the control problem consisting in minimizing $\mathcal{C}^{\lambda}_{\mathrm{KL}}$ w.r.t. the control $\mathrm{u}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ belongs to the KL-control class and has a closed form solution Todorov [2009], Kappen et al. [2012]. The probability distribution of an optimal path $\mathrm{u}^{*}_{\mathrm{KL}}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)$ that minimizes equation (5) is

$$\mathrm{u}^{*}_{\mathrm{KL}}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right) = \frac{\mathrm{p}\left(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t}\right)}{\left\langle \phi(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}})\right\rangle_{\mathrm{p}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t})}}\phi(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}), \tag{6}$$

with

$$\phi(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}) := \exp\left(-\lambda^{-1}\sum_{\mathrm{t}'=\mathrm{t}+1}^{\mathrm{T}} \mathrm{r}(\mathrm{x}_{\mathrm{t}'},\mathrm{t}')\right). \tag{7}$$

Plugging this into equation (5) and minimizing gives the optimal cost-to-go

$$J^{\lambda}_{\mathrm{KL}}(\mathrm{x},\mathrm{t}) = \mathrm{r}(\mathrm{x},\mathrm{t}) - \lambda \log \left\langle \phi(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}})\right\rangle_{\mathrm{p}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t})}, \tag{8}$$

which can be numerically approximated using paths sampled from the uncontrolled dynamics $\mathrm{p}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},\mathrm{t})$.

The optimal control corresponding to equation (4) corresponds to a state transition probability distribution that is obtained by marginalization in equation (6). It is expressed in terms of the uncontrolled transition probability $\mathrm{p}(\mathrm{x}'|\mathrm{x})$ and the (exponentiated) optimal cost-to-go:

$$\mathrm{u}^{*}_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t}) = \sum_{\mathrm{x}_{\mathrm{t}+2:\mathrm{T}}} \mathrm{u}^{*}_{\mathrm{KL}}\left(\mathrm{x}_{\mathrm{t}+1} = \mathrm{x}', \mathrm{x}_{\mathrm{t}+2:\mathrm{T}}|\mathrm{x},\mathrm{t}\right) \propto \mathrm{p}(\mathrm{x}'|\mathrm{x}) \exp\left(-\frac{J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1)}{\lambda}\right). \tag{9}$$

This resembles a Boltzmann distribution with temperature λ where the optimal cost-to-go takes the role of an energy. The effect of the temperature becomes clear: for high values of λ, $\mathrm{u}^{*}_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ deviates only a little from the uncontrolled dynamics $\mathrm{p}(\mathrm{x}'|\mathrm{x})$, thus the optimal control has a weak influence on the system. In contrast, for low values of λ, the exponential in equation (9) becomes very pronounced for the state(s) x' with the smallest cost-to-go $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1)$, suppressing the transition probabilities to suboptimal states x'. Thus the control has a very strong effect on the process. In the limit of λ going to zero, the controlled process becomes deterministic, if $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1)$ is not degenerate (meaning there is a unique state x' which minimizes the optimal cost-to-go). In this case the control is so strong that it overpowers the noise completely.

We thus approximate our original (possibly difficult) control problem as a KL-control problem, parametrized by the temperature λ. The approximated optimal cost-to-go $J(\mathrm{x}',\mathrm{t}+1)$ of equation (4) is replaced by the corresponding optimal cost-to-go of the KL-control problem $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1)$ of equation (9) and used to compute the action selection in the original problem.

3.2 A Tractable Example

We now present a tractable example amenable for exact optimal control computation. This example already belongs to the KL-control class, so no approximation is made. The purpose of this analysis it

to show how different values of the temperature λ may lead to qualitatively different optimal solutions and other interesting phenomena.

Let's consider a tree that grows at discrete time-steps, starting with the root node at time $\mathrm{t} = 0$. We represent the tree at time t as a vector $\mathrm{x_t} = (x_0, x_1, ..., x_\mathrm{t})$, where x_t indicates the label of the parent of the node attached at time t. At every time-step, either the tree remains the same or a new node is attached to it. The root node has label 1 and the label 0 is specially used to indicate that no node was added at a given time-step (it is also the label of the parent of the root node). The nodes are labelled in increasing order as they arrive to the tree, so that at time-step t, for a tree with k nodes, $k \leq t, x_\mathrm{t} = 0, 1, \ldots, k$ corresponds to the parent of node $k + 1$ if a node is added or zero otherwise. Thus, the parent vector at time $\mathrm{t} = 1$ is always $\mathrm{x}_1 = (0, 1)$.

Our example is a finite horizon task of $\mathrm{T} = 10$ time-steps and end-cost only. The end-cost implements two control objectives: it prefers trees of large Wiener index while penalising trees with many nodes (more than five, in this case). The Wiener index is the sum of the lengths of the shortest paths between all nodes in a graph. It is maximal for a chain and minimal for a star.

The uncontrolled process is biased to the root: new nodes choose to link the root with probability $3/5$ and uniformly otherwise. More precisely

$$\mathrm{p}(x_{t+1} = j|\mathrm{x_t}) = \begin{cases} \frac{3}{5} & \text{for } j = 1 \\ \frac{2}{5\|\mathrm{x_t}\|_0} & \text{for } (j = 0) \text{ or } j \in \{2, \ldots, \|\mathrm{x_t}\|_0\} \end{cases} \tag{10}$$

$$\mathrm{r}(\mathrm{x_t}, \mathrm{t}) = \begin{cases} -\texttt{Wiener}(\mathrm{x_t})\delta_{\mathrm{t,T}} & \text{if } \|\mathrm{x_t}\|_0 < 5 \\ \delta_{\mathrm{t,T}} & \text{otherwise} \end{cases} \tag{11}$$

where $\|\mathrm{x}\|_0$ denotes the number of non-zero elements in x and `Wiener` the (normalized) Wiener index.

In this setting, the uncontrolled process p tends to grow trees with more than five nodes with many of them attached to the root node, i.e. with low Wiener index. We want to influence this dynamics so that the target configuration, a chain of five nodes (maximal Wiener index) is more likely to be obtained.

Figure 1 (top) shows the state cost r of the final tree that results from choosing the most probable control (MAP solution) as a function of the temperature λ. The exact solution is calculated using dynamic programming Kappen et al. [2012]. We can differentiate three types of solutions, denoted as A, B and C in the figure.

For low temperatures (region A) the control aims to fulfil both control objectives: to find a small network with maximal Wiener index. The optimal strategy does not add nodes initially and then builds a tree of maximal Wiener index (see inset of initial controls in left column of the figure). This type of control (to wait while the target is far in the future) is reminiscent of the delayed choice mechanism described previously Kappen [2005]. This initial waiting period makes sense because if the chain of length 5 would be grown immediately, then at time 6 the size of 5 is already reached. If now an additional node attaches, then the final cost would be zero. However if one first waits and then grows the chain, an accidental node insertion before time 6 would not be so disastrous (actually it may help), as one can then just wait until time 7 to start growing the rest of the tree. So delaying the decision when to start growing the tree helps compensating accidental events.

For intermediate temperatures (region B), the initial control becomes less extreme, as we observe if we compare the left plots between regions A and B. For $\lambda \approx 0.07$, the solution that builds the tree with maximal Wiener index is no longer optimal, since it deviates too much from the uncontrolled

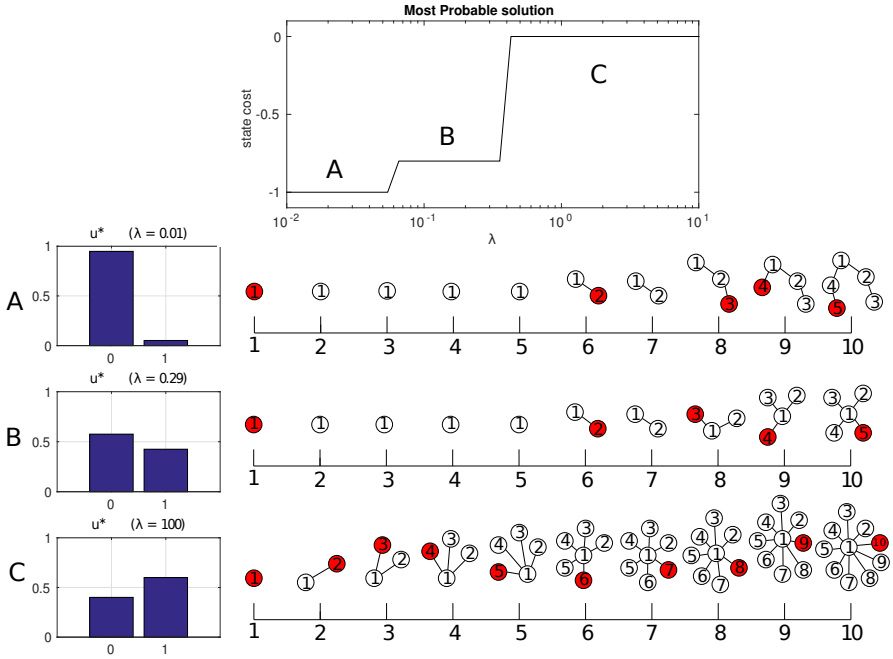

Figure 1: Example of optimal control of tree growth. **(Top)**: the state cost of the most probable solutions as a function of the temperature λ. In region A, the optimal strategy waits until the last time-steps and then grows a tree with maximal Wiener index. In region B, it builds a star of five nodes. Finally, in region C, it follows the uncontrolled dynamics and builds a star of ten nodes. **(Left)**: for each region, the optimal probabilities $\mathrm{u}^*(x_{\mathrm{t}+1}|\mathrm{x_t})$ at $\mathrm{t} = 1$ for the two actions which are initially available: no node addition (0) and adding a node to the root (1). In regions A, B the optimal control favours not adding a new node initially. The sequences on the **right** show how the tree grows. When a new node is added to the tree, it is coloured in red.

dynamics. In region B, the control aims to build a network of five nodes or less, but no longer aims to maximize the Wiener index. The control is characterized by an initial waiting period and the subsequent growth of a tree of five nodes, which are in this case all attached to the root node.

Finally, for high temperatures (region C, $\lambda > 0.4$), the control essentially ignores the cost r and the optimal strategy is to add one node to the root at every time-step, following the uncontrolled process.

From these results we conclude that KL-control as a mechanism for controlling network growth can capture complex phenomena such as transitions between qualitatively different optimal solutions and delayed choice effects.

3.3 Sampling from the KL-optimally controlled dynamics

In this subsection, we explain how we can sample from the optimally controlled dynamics and thereby obtain an estimate of the optimal cost-to-go $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}, \mathrm{t})$ of equation (8).

The probability of an optimally controlled path, equation (6), corresponds to the product of the uncontrolled dynamics by the exponentiated state costs. Hence a naive way to obtain samples from the optimal dynamics, would consist in sampling paths from the uncontrolled dynamics $\mathrm{p}(\mathrm{x}'|\mathrm{x})$ and

weight them by their exponentiated state costs. Using these samples we can then compute expectations from the optimally controlled dynamics. We use that for any function $f(x_{t+1:T})$ we have:

$$\langle f(x_{t+1:T})\rangle_{u^*_{KL}(x_{t+1:T}|x,t)} = \left\langle f(x_{t+1:T})\frac{\phi(x_{t+1:T})}{\langle\phi(x_{t+1:T})\rangle_{p(x_{t+1:T}|x,t)}}\right\rangle_{p(x_{t+1:T}|x,t)}.$$

More precisely, provided a learned model or a simulator of the uncontrolled dynamics $p(x'|x)$, we generate M sample paths $x^{(i)}_{t+1:T}, i = 1,\ldots,M$ from $p(x'|x)$ and compute the weights $\frac{\phi(x^{(i)}_{t+1:T})}{\hat{\phi}}$. The denominator thereby gives with equation (8) an estimate of the optimal cost-to-go as

$$\langle\phi(x_{t+1:T})\rangle_{p(x_{t+1:T}|x,t)} \approx \hat{\phi} := \frac{1}{M}\sum_{i=1}^{M}\phi\left(x^{(i)}_{t+1:T}\right).$$

This method can be combined with resampling techniques Douc and Cappé [2005], Hol et al. [2006] to obtain unweighted samples $x^{\text{opt},(i)}_{t+1:T}$ from the optimal dynamics (for the numerical methods in this article, we have used structural resampling Douc and Cappé [2005], Hol et al. [2006]).

Using such a naive sampling method, however, can be inefficient, specially for low temperatures. While for high temperatures λ basically all weights $\frac{\phi(x^{(i)}_{t+1:T})}{\hat{\phi}}$ are more or less equal, for low temperatures only a few samples with very large weights contribute to the approximation, resulting in very poor estimates.

This is a standard problem in Monte Carlo sampling and can be addressed using the Cross-Entropy (CE) method De Boer et al. [2005], Kappen and Ruiz [2016], which is an adaptive importance sampling algorithm that incrementally updates a baseline sampling policy or sequence of controls. Here we propose to use the CE method in the discrete formulation and use a parametrized Markov process $\widetilde{u}_\omega(x'|x,t)$, with parameters ω, to approximate u^*_{KL}. The CE method in our setting alternates the following steps:

1. In the first step, the optimal control is estimated using M sample paths drawn from a parametrized proposal distribution $\widetilde{u}_\omega(x'|x,t)$.
2. In the second step, the parameters ω are updated so that the proposal distribution becomes closer to the optimal probability distribution.

As a proposal distribution $\widetilde{u}_\omega(x'|x,t)$, we use

$$\widetilde{u}_\omega(x'|x,t) \propto p(x'|x)\exp\left(-\frac{\widetilde{J}_{KL}(x',\omega(t))}{\lambda}\right), \tag{12}$$

which has the same functional form as the optimally controlled transition probabilities in equation (9). The KL-optimal cost-to-go is thereby approximated by a linear sum of time-dependent feature vectors $\psi^t_k(x)$

$$\widetilde{J}_{KL}(x,\omega(t)) = \sum_k \omega_k(t)\psi^t_k(x). \tag{13}$$

The probability distribution of an optimally controlled path, equation (6), can be written as

$$u^*_{KL}\left(x^{(i)}_{t+1:T}|x,t\right) \propto \widetilde{u}_\omega(x^{(i)}_{t+1:T}|x,t)\frac{p\left(x^{(i)}_{t+1:T}|x,t\right)}{\widetilde{u}_\omega(x^{(i)}_{t+1:T}|x,t)}\exp\left(-\lambda^{-1}\sum_{t'=t+1}^{T} r(x^{(i)}_{t'},t')\right). \tag{14}$$

This shows that we can draw samples from the proposal distribution and reweight them with the combined weights

$$\mathrm{w}^{(i)} = \frac{\mathrm{p}(\mathrm{x}^{(i)}_{\mathrm{t+1:T}}|\mathrm{x},\mathrm{t})}{\widetilde{\mathrm{u}}_{\omega}(\mathrm{x}^{(i)}_{\mathrm{t+1:T}}|\mathrm{x},\mathrm{t})}\phi\left(\mathrm{x}^{(i)}_{\mathrm{t+1:T}}\right).$$

The parameters $\omega_k(\mathrm{t})$ of the importance sampler are initialized with zeros, which makes the initial proposal distribution equivalent to the uncontrolled dynamics. The procedure requires the gradients of $\widetilde{\mathrm{u}}_{\omega}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ at each iteration. We describe the details of the CE method in A.

We measure the efficiency of an obtained proposal control using the effective sample size (EffSS), which estimates how many effective samples can be drawn from the optimal distribution. Given M samples with weights $\mathrm{w}^{(i)}$, the EffSS is given by

$$\mathrm{EffSS} = \frac{\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\left(\mathrm{w}^{(i)}\right)^2}{\left(\frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}}\mathrm{w}^{(i)}\right)^2}. \tag{15}$$

If the weights $\mathrm{w}^{(i)}$ are all about the same value, the EffSS is high, indicating that many samples contribute to statistical estimates using the weighted samples. If all weights are equal, the EffSS is equal to the number of samples M. Conversely if the weights $\mathrm{w}^{(i)}$ have a large spread, the EffSS is low, indicating that only few independent samples contribute to statistical estimates. In the extreme case, when one weight is much larger then all others, the EffSS approaches 1.

3.4 Action selection using the KL-approximation

Once we have an estimate of the cost-to-go $J^{\lambda}_{\mathrm{KL}}$, we need to select an action $\mathrm{u} \in \mathcal{U}$ in the original control problem, which is not of the KL-control type. We select the optimal action according to

$$\mathrm{u}^*(\mathrm{x},\mathrm{t}) \approx \mathrm{argmin}_{\mathrm{u}}\left(\mathrm{r}(\mathrm{x},\mathrm{t}) + \left\langle J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1)\right\rangle_{P(\mathrm{x}'|\mathrm{x},\mathrm{u},t)}\right), \tag{16}$$

which requires the computation of $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t+1}},\mathrm{t}+1)$ for every reachable state $\mathrm{x}_{\mathrm{t+1}}$. In growing networks, the number of possible next states (the branching factor) increases quickly, and visiting all of them soon becomes infeasible.

In this subsection we highlight an important benefit of using the KL-approximation as a relaxation of the original problem: the optimally controlled process tends to discard many irrelevant states, specially for small values of λ. This means that $\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ is sparse on x' (only a few next states are relevant for the task), since the cost $J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',t)$ is very large for the corresponding x' where $\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t}) \approx 0$.

Let $\mathrm{x}^{\mathrm{opt}}_{\mathrm{t+1:T}}$ denote a trajectory sampled from the optimally controlled process, as described in the previous section. We compute $\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ using:

$$\hat{\mathrm{u}}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t}) = \left\langle \delta_{\mathrm{x}^{\mathrm{opt}}(\mathrm{t}+1),\mathrm{x}'}\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t+1:T}}|\mathrm{x},\mathrm{t})}, \tag{17}$$

where $\mathrm{x}^{\mathrm{opt}}(\mathrm{t}+1)$ is the first element of the trajectory and $\delta_{\mathrm{x}^{\mathrm{opt}}(\mathrm{t}+1),\mathrm{x}'}$ is the Kronecker delta which is equal one if $\mathrm{x}^{\mathrm{opt}}(\mathrm{t}+1)$ is equal to x', and zero otherwise.

We then compute the optimal cost using equation (9):

$$J^{\lambda}_{\mathrm{KL}}(\mathrm{x}',\mathrm{t}+1) \sim -\log\left(\frac{\hat{\mathrm{u}}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t})}{\mathrm{p}(\mathrm{x}'|\mathrm{x},\mathrm{t})}\right), \tag{18}$$

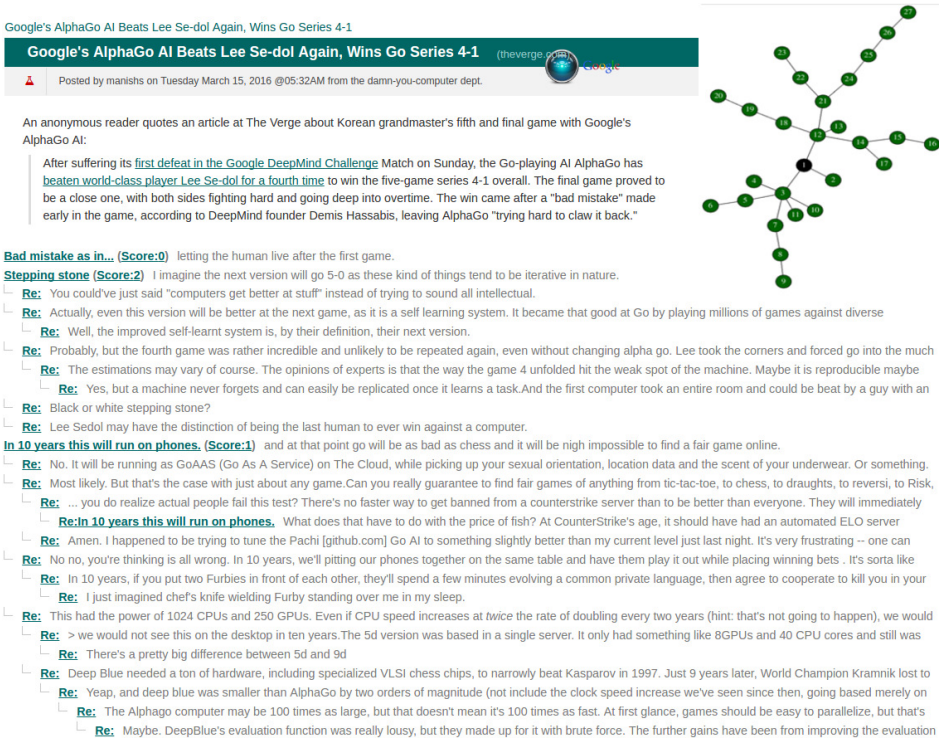

Figure 2: Task illustration: example of an Internet news forum. News are posted periodically and users can write comments either to the original post or to other user's comments, forming a cascade of messages. The figure shows an example of conversation thread taken from Slashdot about *Google's AlphaGo*. The control task is to influence the structure of the conversation thread (shown as a growing tree in the top-right).

where we dropped a term which does not depend on x' and therefore plays no role in the minimization of equation (16). The KL-approximation can help reducing the branching factor because it needs only a few samples to calculate $J^{\lambda}_{\mathrm{KL}}(x', t+1)$ only for the x' where $u^{*}_{\mathrm{KL}}(x'|x,t) > 0$ and thus $J^{\lambda}_{\mathrm{KL}}(x', t)$ has a finite value.

As mentioned earlier, $u^{*}_{\mathrm{KL}}(x'|x,t)$ tends to be more sparse for small values of λ, when the KL-control problem is less noisy. In B we provide analytical details of the two extreme conditions, when λ is zero or infinite, respectively.

4 Application to Conversation Threads

We have described a framework for controlling growing graphs. We now illustrate this framework in the context of growing information cascades. In particular, we focus on the task of controlling the growth of online conversation threads. These are information cascades that occur, for example, in online forums such as weblogs Leskovec et al. [2007], news aggregators Gómez et al. [2008] or the synthesis of articles of Wikipedia Laniado et al. [2011]. In conversation threads, after an initial post appears, different users react writing comments either to the original post or to comments from other users.

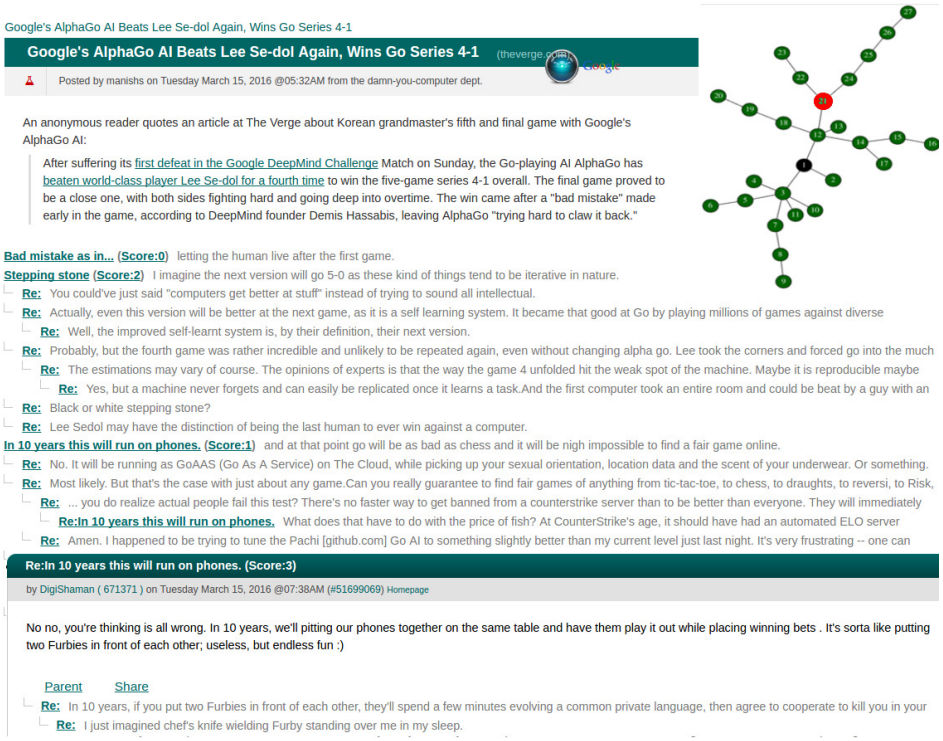

Figure 3: Our proposed control mechanism: in addition to the the threaded conversation, we highlight a comment (red node in the growing tree), suggested to be replied by the user. The choice of suggested comment, shown at the bottom of the page, is calculated using the method described in section 3.3.

Figure 2 shows an example of a conversation thread, taken from Slashdot (`www.slashdot.org`). Users see a conversation thread using a similar hierarchical interface.

The task we consider is to optimize the structure of the generated conversation thread while it grows. The state is thus defined as a growing tree. We assume an underlying (not observed) population of users that keep adding nodes to this tree. Since we can not control directly what is the node that will receive the next comment, we propose the user interface as a control mechanism to influence indirectly the growth process. This can be done in different ways, for example, manipulating the layout of the comments. In our case, the control signal will be to recommend a comment (by highlighting it) to which the next user can reply. Figure 3 illustrates such a mechanism. The action selection strategy introduced in section 3.4 is used to select the comment to highlight. Our goal is thus to modify the structure of a cascade in certain way while it evolves, by influencing its growth indirectly. It is known that the structure of online threads is strongly related with the complexity of the underlying conversation Gómez et al. [2008], Gonzalez-Bailon et al. [2010].

To fully define our control problem, we need to specify the structural cost function, the uncontrolled dynamics, i.e. the equivalent of equations (10) and (11) for this task, and a model of how an action (highlighting a node) changes the dynamics. Globally, this application differs from the toy example of subsection 3.2 in some important ways:

1. The state-space is larger (threads typically receive more than 10 comments).

2. We choose as state-cost function the Hirsch index (h-index), which makes the control task

highly non-trivial.

3. The original problem is not a KL-control problem. We use the action selection method described in section 3.4 to control the growth of the conversation thread.

4.1 Structural Cost Function

We propose to optimize the Hirsch index (h-index) as structural measure. In our context, a cascade with h-index h has h comments each of which have received at least h replies. It is a sensible quantity to optimize, since it measures how distributed the comments of users on previous comments are. A high h-index prevents two extreme cases that occur in a rather poor conversation: the case where a small number of posts attract most of the replies, thus there is no interaction, and the case with deep chains, characteristic of a flame war of little interest for the community. Both cases have a low h-index, while a high h-index spreads the conversation over multiple levels of the cascade.

The h-index is a function of the degree sequence of all nodes in the tree, where the degree of a node is this case is the number of replies plus one, as there is also a link to the parent (replied comment or post). Therefore we use the degree histogram as features $\psi_k^{\mathrm{t}}(\mathrm{x})$ for the parametrized form of the optimal cost-to-go, equation (13). That is, feature $\psi_k^{\mathrm{t}}(\mathrm{x})$ is the number of nodes with degree k in the tree x at time-step t. We model the problem as a finite horizon task with end-cost. Thus, the state cost is defined as $\mathrm{r}(\mathrm{x},\mathrm{t}) = -\delta_{\mathrm{t},\mathrm{T}} \cdot \mathrm{h}(\mathrm{x})$, where $\mathrm{h}(\mathrm{x})$ is the h-index of the tree x.

4.2 Uncontrolled Dynamics for Online Conversation Threads

As uncontrolled dynamics, we use a realistic model that determines the probability of a comment to attract the replies of other users at any time, by means of an interplay between the following features:

- *Popularity* α: number of replies that a comment has already received.
- *Novelty* τ: the elapsed time since the comment appeared in the thread.
- *Root node bias* β: characterizes the level of trendiness of the main post.

Such a model has proven to be successful in capturing the structural properties and the temporal evolution of discussion threads present in very diverse platforms Gómez et al. [2013]. Notice that these features $\theta = (\alpha, \tau, \beta)$ should not be confused with the features $\psi_k^{\mathrm{t}}(\mathrm{x})$ used to encode the cost-to-go.

We represent the conversation thread as a vector of parents $\mathrm{x_t} = (x_0, x_1, ..., x_{\mathrm{t}})$. Given the current state of the thread $\mathrm{x_t}$, the uncontrolled dynamics attaches a new node $\mathrm{t}+1$ to an existing node j with probability

$$\mathrm{p}_\theta(x_{\mathrm{t}+1} = j|\mathrm{x_t}) = \frac{1}{Z_{\mathrm{t}+1}} \left(\mathrm{deg}_{j,\mathrm{t}}\alpha + \delta_{j,1}\beta + \tau^{\mathrm{t}+1-j}\right) \tag{19}$$

with $Z_{\mathrm{t}+1}$ a normalization constant, $\mathrm{deg}_{j,\mathrm{t}}$ the degree of node j at time t and $\delta_{j,1}$ the Kronecker delta function, so parameter β is only nonzero for the root.

Given a dataset composed of S threads $\mathcal{D} := \{\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(\mathrm{S})}\}$ with respective sizes $|\mathrm{x}^{(k)}|$, $k \in \{1, \ldots \mathrm{S}\}$, the parameter vector θ can be learned by minimizing

$$-\log \mathcal{L}(\mathcal{D};\theta) = -\sum_{k=1}^{\mathrm{S}} \sum_{\mathrm{t}=2}^{|\mathrm{x}^{(k)}|} \log \mathrm{p}_\theta(x_{\mathrm{t}+1}^{(k)}|\mathrm{x}_{\mathrm{t}}^{(k)}).$$

We learn the parameters using the Slashdot dataset, which consists of $S = 9,820$ threads, containing more than $2 \cdot 10^6$ comments among $93,638$ users. In Slashdot, the most relevant feature is the preferential attachment, as detailed in Gómez et al. [2013]. This will have implications in the optimal control solution, as we show later.

4.3 Control interaction

The control interaction is done by highlighting a single comment of the conversation. We assume a behavioural model for the user inspired by Craswell et al. [2008], where the user looks at the highlighted comment and decides to reply or not. For simplicity, we assume that the user chooses the highlighted comment with a fixed probability $p' = \alpha/(1+\alpha)$ and with probability $1 - p'$ she chooses to ignore it. If the highlighting of the comment is ignored, the thread grows according to the uncontrolled process. Therefore, α parametrizes the strength of the influence the controller has on the user. For $\alpha \to \infty$, we can fully control the behaviour and for $\alpha = 0$, the thread evolves according to the uncontrolled process. A typical control would have a small α as usually the influence of an controlling agent on a social systems is weak.

4.4 Experimental Setup

To evaluate the proposed framework we use a simulated environment, without real users. We consider a finite horizon task with $T = 50$ with the goal to maximize the h-index at end-time, starting from a thread with a single node as initial condition. The state-space consists of $50! \approx 3^{64}$ states. The thread grows in discrete time-steps. At each time-step, a new node is added to the thread by a (simulated) user. For that, we first choose which node to highlight (optimal action) as described in section 3.4 using equation (16). We then simulate the user as described in section 4.3, so the highlighted node is selected with probability $p' = \alpha/(1+\alpha)$ as the parent of the new node. Otherwise, with probability $1 - p'$, the user ignores the highlighted node and the parent of the new node is chosen according to the Slashdot model, equation (19). This is repeated until the end time.

4.5 Experimental Results

We first analyse the performance of the adaptive importance sampling algorithm described in section 3.3 for different fixed values of λ.

Figure 4 shows the effective sample size (EffSS), equation (15) as a function of the number of iterations of the CE method. We observe that the EffSS increases to reach a stable value. As expected, large temperature (easier) problems result in higher values of EffSS. We can also see that, even for hard problems with low temperature, the obtained EffSS is significantly larger than zero, which allows us to compute the KL-optimal control. In general, the curves are less smooth for smaller values of λ, because a few qualitatively better samples dominate the EffSS, resulting in higher variance. On the other hand we also observe that the EffSS never reaches 100%. This is expected, as this would mean that our parametrized importance sampler perfectly resembles the optimal control, and this is not possible due to the approximation error introduced by the use of features.

We can better understand the learned control by analysing the linear coefficients of the parametrized optimal cost-to-go, equation (13), for this problem. Figure 5 shows the feature weights $\omega_k(t)$, at different times $t = 1, \dots, T$, after convergence of the CE method. Feature k corresponds to the number of nodes with degree k in the tree, after a new node arrives. The parent node to which the new node

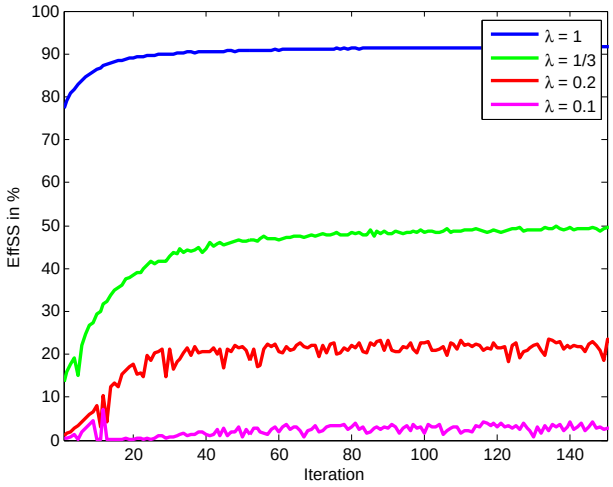

Figure 4: **Evaluation of the inference step**: The Effective sampling size (EffSS) increases after several iterations of the cross-entropy method. As expected, large values of the temperature λ result in higher values of EffSS. We use $M = 10^5$ samples to compute the EffSS. The EffSS is measured here in percent of the maximum number of samples M.

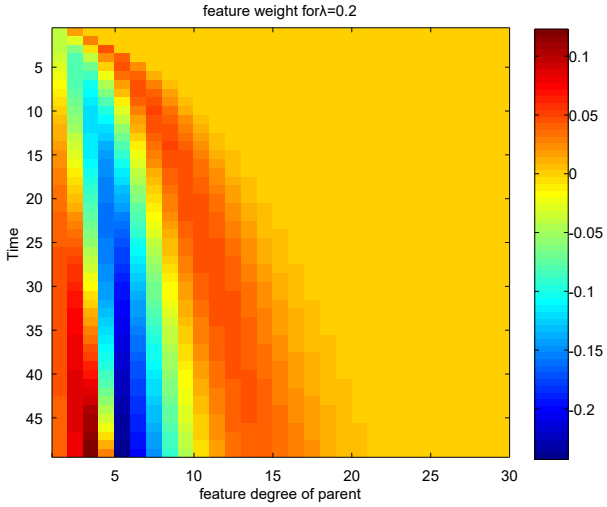

Figure 5: **The learned importance sampler**: The figure shows the time-dependent parameters of the learned expected cost-to-go for $\lambda = 0.2$. Each pixel is the parameter of a feature at a certain time. The features are the degrees of the parent node after the new child attaches. The colour represents the weight of the parameter. Large negative weights (pixels in blue colour) stand for a low cost and thus a desirable state, while large positive weights (red pixels) stand for high cost and thus undesirable states. At all times there is a desirable degree which the parent should have and higher as well as lower degrees are inhibited. This desirable degree is small at early times and becomes larger at later times.

has attached is thereby the only node whose degree changes (the degree increases by 1). Thus a high weight for a feature which measures the number of nodes with a certain degree k results in a low probability of attaching to a node with degree $k-1$. Conversely low, or large negative weights thus correspond to nodes which have a high probability of becoming the parent of the next node which is added. We observe that there is an intermediate preferred degree (large negative weight, in blue). This is the preferred degree of the parent of the new node, and this preferred degree increases with time, reaching a value of 5 at t $= 50$.

Does this strategy make sense? The maximum h-index of a tree of 50 nodes is 7, and it is achieved if 6 nodes have exactly 7 children and one node has 8. However, achieving such a configuration requires a very precise control. For example, increasing too much the degree of a node, say up to 9, prevents the maximum h-index to be reached, as there are not enough links left, due to the finite horizon. Thus, in this setting, steering for the maximal possible h-index is not optimal. The controller prefers all parents to have a degree of 5 and not less, but also not much more. As having more than five parents with degree at least five will result in an h-index of 5 we conclude that the control seems to aim for a target h-index of 5, while preventing *wasting* links to higher or lower degree nodes, which would not contribute to achieve that target.

The interpretation of why the preferred degree increases with time involves the uncontrolled dynamics. Remember that the most relevant term in equation (19) for the considered dataset corresponds to the preferential attachment, parametrized by α. This term boosts high-degree nodes to get more links. If this happens, most of the links end up attached to a few parents, and this effect can only be suppressed by a strong control. The controller prevents that self-amplifying effect by aiming initially for an overall low degree, preventing a high impact of the preferential attachment. This keeps the process controllable and allows for a more equal distribution of the links.

After having evaluated the sampling algorithm, we evaluate the proposed mechanism for actual control of the conversation thread. As described in section 3.4, in our simulated scenario, we highlight the node as the parent which minimizes the computed expected cost-to-go.

Figure 6 shows the evolution of the h-index using different control mechanisms. The blue curve shows how the h-index changes under the uncontrolled dynamics. On average, it reaches a maximum of about 3.7 after 50 time steps. In green, we show the evolution of the h-index under a KL-optimal controlled case, for temperature $\lambda = 0.2$. As expected, we observe a faster increase, on average, than using the uncontrolled dynamics. The maximum is about 4.7.

The red and black curves show the evolution of the h-index using the control mechanism described in subsections 3.4 and 4.3, where we select actions using the expected cost-to-go $J^{\lambda}_{\mathrm{KL}}$ of the KL-optimal control with $\lambda = 0.2$, for $\alpha = 1$ and $\alpha = 0.5$, respectively. In both cases the obtained h-index is even higher than the one obtained with the KL-control relaxation. Therefore, the objective for this task, to increase the h-index, can be achieved through our action selection strategy. As expected, a stronger interaction strength $\alpha = 1$ leads to higher h-indices than a lower strength $\alpha = 0.5$.

Finally, in Figure 7 we show examples of a real discussion thread from the dataset (Slashdot), a thread generated from the learned model (uncontrolled process) and one resulting from applying our action selection strategy. The latter has higher h-index.

5 Discussion

We have addressed the problem of controlling the growth process of a network using stochastic optimal control with the objective to optimize a structural cost that depends on the topology of the

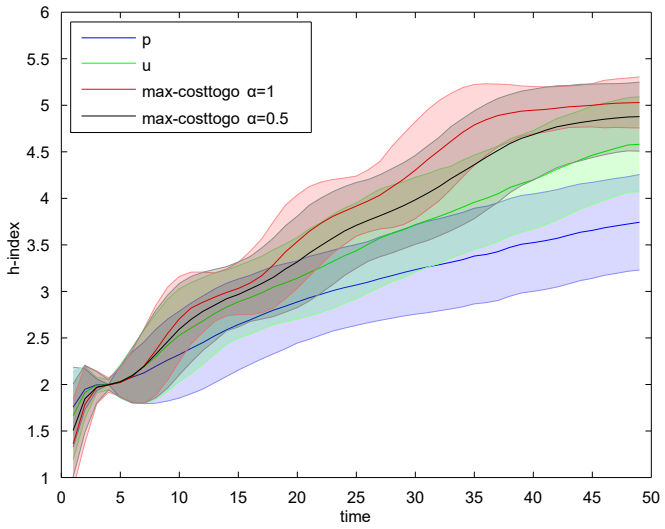

Figure 6: **Evaluation of the actual control**: uncontrolled dynamics (blue), KL-optimally controlled dynamics (green) action selection based control for $\alpha = 1$ (red) and $\alpha = 0.5$ (black). The KL-optimally controlled dynamics, which optimize the sum of the λ-weighted KL-term and the end cost, shifts the final mean value from about 3.7 to about 4.7. The action selection based control, which is aiming to optimize the end cost only, is able to shift the h-index to even higher values then the KL-optimal control. For the controlled dynamics, $\lambda = 0.2$ for all three cases. To compute the control in each time-step we sample 1000 trajectories. The statistics where computed using 1000 samples for each of the three cases.

growing network. The main difficulty of such a problem is the exploding size of the state space, which grows (super-)exponentially with the number of nodes in the network and renders exact dynamic programming infeasible.

We have shown that a convenient way to address this problem is using KL-control, where a regularizer is introduced which penalizes deviations from the natural network growth process. One advantage of this approach is that the optimal control can be solved by sampling. The difficulty of the sampling is controlled by the strength of the regularization, which is parametrized by a temperature parameter λ: for high temperatures the sampling is easy, while for low temperatures, it becomes hard. This is in contrast to standard dynamic programming, whose complexity is directly determined by the number of states and independent of λ.

In order to tackle the more challenging low temperature case, we have introduced a feature-based parametrized importance sampler and used adaptive importance sampling for optimizing its parameters. This allows us to sample efficiently in the low temperature regime. For control problems which cannot directly be formulated as KL-control problems, we have proposed to use the solution of a related KL-control problem as a proxy to estimate the effective values of possible next network states. These expected effective values are subsequently used in a greedy strategy for action selection in the original control problem. This action selection mechanism benefits from the sparsity induced by the optimal KL-control solution.

We have illustrated the effectiveness of our method on the task of influencing the growth of

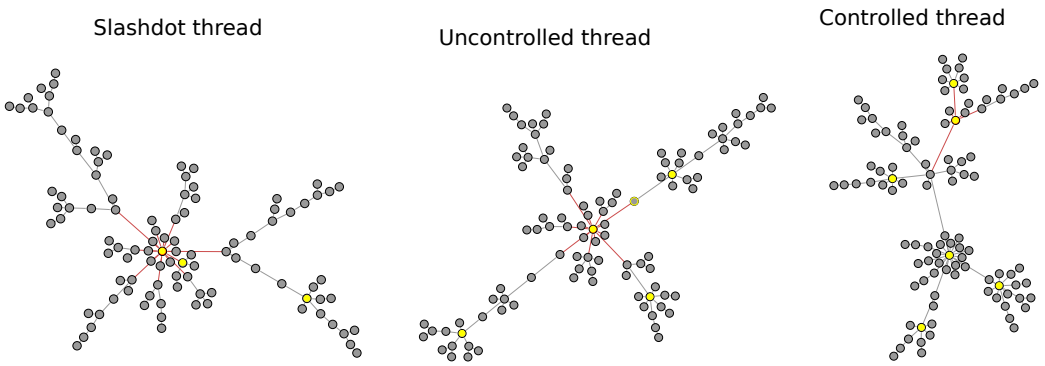

Figure 7: Examples of threads. A thread from the data (Slashdot), an uncontrolled thread generated from the model and a controlled thread. The nodes that contribute to the h-index are coloured in yellow. The h-index for the data and the uncontrolled thread is 4 and 6 for the controlled one.

conversation cascades. Our control seeks to optimize the structure of the cascade, as it evolves in time, to maximize the h-index at a final time. This task is non-trivial and characterized by a sparse, delayed reward, since the h-index remains constant during most of the time, and therefore a greedy strategy is not possible.

Our approach for controlling network growth is inspired in recent approaches to optimal decision-making with information-processing constraints Todorov [2009], Tishby and Polani [2011], Kappen et al. [2012], Theodorou and Todorov [2012], Rawlik et al. [2012]. The Cross-Entropy method has been explored previously in the continuous case Kappen and Ruiz [2016]. The continuous formulation of this class of problems has been used in robotics, using parametrized policies Theodorou et al. [2010], Levine and Koltun [2013], Gómez et al. [2014]. In economics, the question of altering social network structure in order to optimize utility has been addressed mainly from a game theoretical point of view, under the name of strategic network formation Jackson and Watts [2002], Bloch and Jackson [2007]. To the best of our knowledge, the problem of network formation has not yet been addressed from a stochastic optimal control perspective.

The standard approach to address the problem of controlling a complex, networked system is to directly try to control the dynamics *on* the network Liu et al. [2011], Cornelius et al. [2013]. This approach considers the classical notion of structural controllability as the capability of being driven from any initial state to any desired final state within finite time. Optimal control in thus referred to the situation where a network can be fully controlled using only one driving signal. This idea is also prevalent in the influence maximization problem in social networks Kempe et al. [2003], Farajtabar et al. [2014, 2015], which consists in finding the subset of driver (most influential) nodes in a network.

Since the controllability of the dynamics *on* the network depends crucially on the topology, several works have considered the idea of changing the network structure is some way that favours structural controllability.

For example, the perturbation approach introduced in Wang et al. [2012] looks for the minimum

number of links that needs to be added so that the perturbed network can be fully controlled using a single input signal. In Hou et al. [2015], a method to enhance structural controllability of a directed network by changing the direction of a small fraction of links is proposed. More recently, Wang et al. [2016] analyzed node augmentation of directed networks while insisting that the minimum number of drivers remains unchanged.

The main difference between our approach and these approaches is that, rather than considering the controllability of the dynamical system *on* the underlying network, our optimal control task is defined on *the structure* of the network itself, regardless of the dynamical system defined on it. In some sense, our results complement these approaches. For example, one could use our optimal control approach to shape the growth of the network in a way that the structural controllability, understood as the state cost function, is optimized.

Acknowledgements

This project is co-financed by the Marie Curie FP7-PEOPLE-2012-COFUND Action, Grant agreement no: 600387, the Marie Curie Initial Training Network NETT, project N. 289146 and the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502).

A Adaptive Importance Sampling for KL-Optimal Control Computation using the Cross-Entropy method

Here we show how the time-dependent weights $\omega_k(\mathrm{t})$ of the importance sampler are updated such that $\widetilde{\mathrm{u}}_\omega(\mathrm{x}'|\mathrm{x},\mathrm{t})$ becomes closer to the optimal sampling distribution. This corresponds to the second step of the Cross-Entropy method described in subsection 3.3. For clarity in the derivations, we will replace $\mathrm{p}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)$ and $\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)$ by p and $\mathrm{u}^*_{\mathrm{KL}}$, respectively, in the expectations. The closeness of the two distributions $\widetilde{\mathrm{u}}_\omega(\mathrm{x}'|\mathrm{x},\mathrm{t})$ and $\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}'|\mathrm{x},\mathrm{t})$ can be measured as the cross entropy between the path $\mathrm{x}_{1:\mathrm{T}}$ probabilities under these two Markov processes:

$$\begin{aligned}\mathrm{KL}\left[\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\parallel\widetilde{\mathrm{u}}_\omega(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\right] &= \left\langle\log\frac{\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}{\widetilde{\mathrm{u}}_\omega(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}}\\ &= -\left\langle\log\widetilde{\mathrm{u}}_\omega(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} + const. =: -\mathrm{D}(\omega), \qquad (20)\end{aligned}$$

where the constant term $\left\langle\log\mathrm{u}^*_{\mathrm{KL}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}}$ is dropped.

We minimize equation (20) by gradient descent. At iteration l, the gradient $\mathrm{D}(\omega^{(l)})$ with respect to $\omega_k(\mathrm{t})$ is given by

$$\frac{\partial\mathrm{D}(\omega^{(l)})}{\partial\omega_k(\mathrm{t})} = -\left\langle\frac{\partial}{\partial\omega_k(\mathrm{t})}\log\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}}$$

where

$$\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0) = \frac{1}{\mathrm{Z}}\mathrm{p}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)\prod_{\mathrm{t}=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}+1},\omega(\mathrm{t}))}{\lambda}\right)$$
$$\mathrm{Z} = \left\langle\prod_{\mathrm{t}'=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda}\right)\right\rangle_{\mathrm{p}}$$

with the normalization constant Z. This leads to

$$\frac{\partial \mathrm{D}(\omega^{(l)})}{\partial\omega_k(\mathrm{t})} = -\left\langle\frac{\partial}{\partial\omega_k(\mathrm{t})}\left(\log\mathrm{p}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0) - \sum_{\mathrm{t}'=0}^{\mathrm{T}-1}\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda} - \log\mathrm{Z}\right)\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} \tag{21}$$

where we can drop the first term as it is independent of $\omega(\mathrm{t})$. The second term can be evaluated using the definition of $\widetilde{J}_{\mathrm{KL}}$, equation (13).

Further, plugging in Z we get

$$\begin{aligned}
\frac{\partial \mathrm{D}(\omega^{(l)})}{\partial\omega_k(\mathrm{t})} &= \lambda^{-1}\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} + \frac{\partial}{\partial\omega_k(\mathrm{t})}\left\langle\log\left\langle\prod_{\mathrm{t}'=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda}\right)\right\rangle_{\mathrm{p}}\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} \\
&= \lambda^{-1}\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} + \frac{\partial}{\partial\omega_k(\mathrm{t})}\log\left\langle\prod_{\mathrm{t}'=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda}\right)\right\rangle_{\mathrm{p}} \\
&= \lambda^{-1}\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} + \frac{1}{\mathrm{Z}}\frac{\partial}{\partial\omega_k(\mathrm{t})}\left\langle\prod_{\mathrm{t}'=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda}\right)\right\rangle_{\mathrm{p}} \\
&= \lambda^{-1}\left(\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} - \frac{1}{\mathrm{Z}}\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\prod_{\mathrm{t}'=0}^{\mathrm{T}-1}\exp\left(-\frac{\widetilde{J}_{\mathrm{KL}}(\mathrm{x}_{\mathrm{t}'+1},\omega(\mathrm{t}'))}{\lambda}\right)\right\rangle_{\mathrm{p}}\right) \\
&= \lambda^{-1}\left(\left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\mathrm{u}^*_{\mathrm{KL}}} - \left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}\right) \\
&= \lambda^{-1}\left(\frac{\left\langle\frac{\mathrm{p}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}{\widetilde{\mathrm{u}}_{\omega}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}\phi\left(\mathrm{x}_{1:\mathrm{T}}\right)\left(\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right)\right\rangle_{\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}}{\left\langle\frac{\mathrm{p}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}{\widetilde{\mathrm{u}}_{\omega}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}\phi\left(\mathrm{x}_{1:\mathrm{T}}\right)\right\rangle_{\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}} - \left\langle\psi^{\mathrm{t}}_k(\mathrm{x}_{\mathrm{t}+1})\right\rangle_{\widetilde{\mathrm{u}}_{\omega^{(l)}}(\mathrm{x}_{1:\mathrm{T}}|\mathrm{x},0)}\right),
\end{aligned} \tag{22}$$

where we have used the estimates from the importance sampling step and equation (9).

The update rule for the parameters becomes

$$\omega_k^{(l+1)}(\mathrm{t}) = \omega_k^{(l)}(\mathrm{t}) + \eta\frac{\partial \mathrm{D}(\omega^{(l)})}{\partial\omega_k(\mathrm{t})}, \tag{23}$$

for some learning rate η. Algorithm 1 summarizes the CE method applied to this context.

Algorithm 1 Cross-Entropy Method for KL-control

Require: importance sampler $\widetilde{\mathrm{u}}_\omega$,
feature space $\psi(\cdot)$,
number of samples M,
learning rate η

$l \leftarrow 0$
$\omega_k^{(l)}(\mathrm{t}) \leftarrow 0$, Initialize weights for all k, t, l
$\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}^{(i)} \leftarrow$ draw M sample trajectories $\sim \widetilde{\mathrm{u}}_{\omega^{(l)}}$, $i = 1, \ldots, \mathrm{M}$
repeat
  compute gradient $\frac{\partial \mathrm{D}(\omega^{(l)})}{\partial \omega_k(\mathrm{t})}$ using equation (21)
  $\omega_k^{(l+1)}(\mathrm{t}) \leftarrow \omega_k^{(l)}(\mathrm{t}) + \eta \frac{\partial \mathrm{D}(\omega^{(l)})}{\partial \omega_k(\mathrm{t})}$ for all k, t, l
  $\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}^{(i)} \leftarrow$ draw M samples $\sim \widetilde{\mathrm{u}}_{\omega^{(l+1)}}$
  $l \leftarrow l + 1$
until convergence

B Analyzing the KL-optimal cost-to-go based action selection

We have introduced an action selection framework which is based on an approximation of the optimal cost-to-go $J(\mathrm{x}', \mathrm{t})$ by the optimal cost-to-go $J_{\mathrm{KL}}^{\lambda}(\mathrm{x}', \mathrm{t}+1)$ of a parametrized family of KL-control problems which share the same state cost $\mathrm{r}(\mathrm{x}, \mathrm{t})$.

Why is this a good idea? Consider the two extreme cases where the temperature λ, which parametrizes the family of equivalent KL-control problems, is zero or infinite, respectively.

Extreme case $\lambda \to 0$ (zero temperature): The total cost in the KL-control problem becomes equal to the total cost in the original control problem, equation (1), as the KL term vanishes. The KL-optimal control becomes deterministic:

$$\lim_{\lambda \to 0} \mathrm{u}_{\mathrm{KL}}^*(\mathrm{x}'|\mathrm{x}, \mathrm{t}) = \lim_{\lambda \to 0} \frac{\mathrm{p}(\mathrm{x}'|\mathrm{x}) \exp\left(-\frac{J_{\mathrm{KL}}^{\lambda}(\mathrm{x}', \mathrm{t}+1)}{\lambda}\right)}{Z} = \begin{cases} 1 & \text{for } \mathrm{x}' = \operatorname{argmin} J_{\mathrm{KL}}^{\lambda}(\mathrm{x}', \mathrm{t}+1) \\ 0 & \text{otherwise} \end{cases}, \quad (24)$$

where Z is a normalization constant.

Thus, for $\lambda \to 0$, the KL-control problem becomes identical to the original problem if the system is fully controllable, i.e. for every t, x and $\tilde{\mathrm{x}}$ there is a $\mathrm{u}_{\tilde{\mathrm{x}}} \in \mathcal{U}$ such that $\mathrm{p}(\mathrm{x}'|\mathrm{x}, \mathrm{t}, \mathrm{u}_{\tilde{\mathrm{x}}}) = \delta_{\tilde{\mathrm{x}}, \mathrm{x}'}$.

Extreme case $\lambda \to \infty$ (infinite temperature): For this case, using equation (8) we get

$$\begin{aligned} J_{\mathrm{KL}}^{\infty}(\mathrm{x}, \mathrm{t}) &= \lim_{\lambda \to \infty} J_{\mathrm{KL}}^{\lambda}(\mathrm{x}, \mathrm{t}) \\ &= \mathrm{r}(\mathrm{x}, \mathrm{t}) - \lim_{\lambda \to \infty} \lambda \log \left(\left\langle \exp\left(-\lambda^{-1} \sum_{\mathrm{t}'=\mathrm{t}+1}^{\mathrm{T}} \mathrm{r}(\mathrm{x}_{\mathrm{t}'}, \mathrm{t}') \right) \right\rangle_{\mathrm{p}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x}, \mathrm{t})} \right) \\ &= \mathrm{r}(\mathrm{x}, \mathrm{t}) + \left\langle \sum_{\mathrm{t}'=\mathrm{t}+1}^{\mathrm{T}} \mathrm{r}(\mathrm{x}_{\mathrm{t}'}, \mathrm{t}') \right\rangle_{\mathrm{p}(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x}, \mathrm{t})}. \end{aligned}$$

Using equation (1) and the definition of the uncontrolled dynamics, we can write

$$J^{\infty}_{\mathrm{KL}}(\mathrm{x},\mathrm{t}) = \mathrm{r}\left(\mathrm{x},\mathrm{t}\right) + \left\langle \sum_{\mathrm{t}'=\mathrm{t}+1}^{\mathrm{T}} \mathrm{r}\left(\mathrm{x}_{\mathrm{t}'},\mathrm{t}'\right) \right\rangle_{P(\mathrm{x}_{\mathrm{t}+1:\mathrm{T}}|\mathrm{x},0,\mathrm{t})} = \mathcal{C}\left(\mathrm{x},\mathrm{t},0\right). \tag{25}$$

Thus, for $\lambda \to \infty$, the KL-optimal cost-to-go becomes equal to the total cost in the original control problem under the uncontrolled dynamics (using $\mathrm{u} = 0$). Having this equation (16) can be written as

$$\mathrm{u}^{*}(\mathrm{x},\mathrm{t}) \approx \mathrm{argmin}_{\mathrm{u}} \left(\mathrm{r}(\mathrm{x},\mathrm{t}) + \left\langle \mathcal{C}\left(\mathrm{x}',\mathrm{t}+1,0\right)\right\rangle_{P(\mathrm{x}'|\mathrm{x},\mathrm{u},\mathrm{t})} \right). \tag{26}$$

In this case, the action selection is equivalent to optimize an expected total cost assuming the system will evolve according to the free dynamics in the future. Thus the infinite temperature control can be used if one wants to guarantee that the obtained solution will not be worse than the solution obtained with zero control. Choosing a lower λ, however, might in practice work better (as we also have shown in section 4) but has no theoretical guarantee.

We can conclude that our action selection strategy is meaningful in the two extreme cases, $\lambda \to \infty$ and $\lambda \to 0$. Also this analysis suggests that, if the available set of actions $\mathrm{u} \in \mathcal{U}$ offers a strong control over the system dynamics, it is more convenient to use a $J^{\lambda}_{\mathrm{KL}}$ with a low temperature λ.

References

H. Amini, R. Cont, and A. Minca. Resilience to contagion in financial networks. *Mathematical finance*, 26(2):329–365, 2016.

D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.

F. Bloch and M. O. Jackson. The formation of networks with transfers among players. *Journal of Economic Theory*, 133(1):83–110, 2007.

D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996): 1194–1197, 2010.

D. Centola and A. Baronchelli. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7):1989–1994, 2015. doi: 10.1073/pnas.1418838112.

S. P. Cornelius, W. L. Kath, and A. E. Motter. Realistic control of network dynamics. *Nature Communications*, 4(1942), Jun 2013. doi: 10.1038/ncomms2939.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM, 2008.

R. Dai and M. Mesbahi. Optimal topology design for dynamic networks. In *Decision and Control and European Control Conference, 2011 50th IEEE Conference on*, pages 1280–1285, 2011.

P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE, 2005.

V. M. Eguíluz and K. Klemm. Epidemic threshold in structured scale-free networks. *Physical Review Letters*, 89(10):108701, Aug 2002.

M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In *Advances in neural information processing systems*, pages 2474–2482, 2014.

M. Farajtabar, Y. Wang, M. Rodriguez, S. Li, H. Zha, and L. Song. COEVOLVE: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, pages 1945–1953, 2015.

P. Gai and S. Kapadia. Contagion in financial networks. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 2010. ISSN 1364-5021. doi:10.1098/rspa.2009.0410.

J. Gao, Y.-Y. Liu, R. M. D'Souza, and A.-L. Barabási. Target control of complex networks. *Nature communications*, 5(5415), 2014.

P. Giudici and A. Spelta. Graphical network models for international financial flows. *Journal of Business & Economic Statistics*, 34(1):128–138, 2016.

S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2015.

V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in Slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654. ACM, 2008.

V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, 16(5-6):645–675, 2013. ISSN 1386-145X. doi: 10.1007/s11280-012-0162-8.

V. Gómez, H. J. Kappen, J. Peters, and G. Neumann. Policy search for Path-Integral control. In *Machine Learning and Knowledge Discovery in Databases*, pages 482–497. Springer, 2014.

S. Gonzalez-Bailon, A. Kaltenbrunner, and R. E. Banchs. The structure of political discussion networks: a model for the analysis of online deliberation. *Journal of Information Technology*, 25(2): 230–243, 2010.

J. D. Hol, T. B. Schon, and F. Gustafsson. On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop*, pages 79–82. IEEE, 2006.

L. Hou, S. Lao, M. Small, and Y. Xiao. Enhancing complex network controllability by minimum link direction reversal. *Physics Letters A*, 379(20):1321–1325, 2015.

M. O. Jackson and A. Watts. The evolution of social and economic networks. *Journal of Economic Theory*, 106(2):265–295, 2002.

H. J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95 (20):200201, 2005. doi: 10.1103/PhysRevLett.95.200201.

H. J. Kappen and H. C. Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 162(5):1244–1266, 2016. ISSN 1572-9613. doi: 10.1007/s10955-016-1446-7.

H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, 2012.

D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: network and tree structure of wikipedia discussion pages. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 177–184, 2011.

J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SIAM International Conference on Data Mining*, volume 7, pages 551–556, 2007.

S. Levine and V. Koltun. Guided policy search. *Proceedings of The 30th International Conference on Machine Learning*, 3:1–9, 2013.

Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. Controllability of complex networks. *Nature*, 473(7346): 167–173, 2011.

B. Mohar and T. Pisanski. How to compute the Wiener index of a graph. *Journal of Mathematical Chemistry*, 2(3):267–277, 1988. ISSN 0259-9791. doi: 10.1007/BF01167206.

R. Olfati-Saber, A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, Apr 2001. doi: 10.1103/PhysRevLett.86.3200.

K. Rawlik, M. Toussaint, and S. Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Int. Conf. on Robotics Science and Systems (R:SS 2012)*, 2012.

E. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and KL control. In *Decision and Control, 2012 IEEE 51st Annual Conference on*, pages 1466–1473. IEEE, 2012.

E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, 11(Nov):3137–3181, 2010.

N. Tishby and D. Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

J. Wang, X. Yu, and L. Stone. Effective augmentation of complex networks. *Scientific Reports*, 6: 25627, 2016.

W.-X. Wang, X. Ni, Y.-C. Lai, and C. Grebogi. Optimizing controllability of complex networks by minimum structural perturbations. *Physical Review E*, 85(2):026115, Feb 2012. doi: 10.1103/PhysRevE.85.026115.

G. Yan, G. Tsekenis, B. Barzel, J. J. Slotine, Y. Y. Liu, and A. L. Barabasi. Spectrum of controlling and observing complex networks. *Nature Physics*, 11(9):779–786, 2015.