Optimizing the Constrained Estimate of Random Walks

RONG ZHOU\textsuperscript{1}, YONG ZHANG\textsuperscript{1}, AND LI NING\textsuperscript{2},
\textsuperscript{1}Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen, China (e-mail: rong.zhou, zhangyong@siat.ac.cn)
\textsuperscript{2}Maketran (Shenzhen) Technology Ltd. (e-mail: li.ning@siat.ac.cn)

Corresponding author: Li Ning (e-mail: li.ning@siat.ac.cn).

This work was supported in part by NSFC grant number 61433012, U1435215; Shenzhen basic research grant number JCYJ20160229195940462.

ABSTRACT We introduce the problem of optimizing the constrained estimate of random walks on the probability networks (which are formally defined as the weighted directed graphs in which the total outgoing weight of any node is at most 1). In this problem, it is required to find the size-constrained set of nodes, which maximized the hitting probability (before the time limit) of the randomly initialized random walk. The problem is proved NP-Hard, and we propose an algorithm with polynomial time complexity and constant approximation ratio.

INDEX TERMS Network theory (graphs), Computational complexity, Optimization

I. INTRODUCTION

As early as 1905, Karl Pearson proposed the concept of random walk, as a simple model to describe the mosquitoes movement in the forest [9]. In the following decades, plenty of studies have been investigated the knowledge about random walks. Aside from the theoretical efforts devoted to the mathematical proofs, the idea of random walk has been found useful in applications of many practical fields, including statistical physics, information diffusion, financial analysis, and biological systems.

For example, consider a microscopic system, which is a primary subject in the studies of statistical physics. The system changes randomly from state to state. Due to the dynamic nature of the system, it is difficult for the observer to know exactly the current state of the system since it keeps changing and might be quite different from the state at which the observer measured. In such situation, the best thing to do is to estimate the current state based on the recently measuring result and the knowledge of the considered system. Formally speaking, when the observer estimates the system’s state, he may want to determine the probability with which the system’s state becomes $s'$ at time $t'$, given the fact that the system was at state $s$ at time $t$. In some cases, the observer does not care about the exact state at time $t'$. Instead, he may want to know whether the system has reached state $s'$ by time $t'$. This is about estimating the system’s state in a time period, and it could be much more difficult than estimating the state at an exact time point, as demonstrated in this paper.

Opinion influence in social networks is another application where the random walk technique is useful. The influence process can be imagined as an opinion propagating from one individual to other related individuals and going on to further spread. However, it should be noticed that in practical, the influence processes are often progressive, which means once an individual is influenced by an opinion, it won’t change unless another opinion wins the place. On the other hand, a random walk is usually non-progressive, since the states it switches between are in general exclusive. Actually, the links between the influence processes and the random walks are in the coupling form, as revealed in the later argument.

A. OUR RESULTS

In this work, we first introduce the problem of optimizing the constrained estimate of random walks on a given weighted graph. This problem is proved NP-Hard, due to a reduction to the problem of vertex cover, as shown in Section III. Then in Section IV we reveal that there is a polynomial-time approach which can approximate the maximum constrained estimate with a constant ratio. In Section V, we investigate the influence processes and prove the coupling relation to the random walks. In the discussions, we consider a strict version of the estimate optimization and show that it is much easier to solve. We also discuss the situations when the initial state of the random walk is selected with bias involved, and show...
that with a minor generalization, the proposed approach can still be applied.

**B. RELATED WORKS**

Ever since Karl Pearson proposed the idea [9], random walks have been extensively studied in the literature. Readers can refer to [7] for the fundamental knowledge and the well-studied topics. An introduction to the typical applications using random walks can be found in [10].

In [11], Hunt studied the problem to identifying optimal spreaders with the random walk model, in which the target is to find the spreader (a set of nodes) that minimizes the first hitting time of random walks starts outside the spreader.

The topic of influences on networks has been extensively studied in recent two decades. As the rising of social networking applications, researchers get more opportunities to reach the large and complex network data which reflect the social interconnections. Modeling the influences between people provides useful supports to spread the ideas or innovations in the more efficient way, and hence it has great potential in the business view. Inspired by Domingos and Richard’s work about viral marketing [2], Kempe et al. considered the problem of maximizing the influences through social networks, and proposed algorithms with constant approximation ratio for the linear threshold model and the independent cascade model [6].

The connection between the influence process and the random walk has also attracted researchers’ attention. In [5], Holley et al. revealed the dual relation between the random walks and the voter model, which is close to the one we are going to investigate in the later argument. Asavathiratham made a similar analysis for the binary influence model [1], and Even-Dar and Shapira studied the influence maximization problem of the voter model [3].

Note that it is still possible to capture the influence behavior with classical random walks. In [12], Zhao et al. designed a random-walk-based measure of the influence and studied the maximization problem.

**II. PRELIMINARIES**

**A. PROBABILITY NETWORK**

A probability network is a non-negative weighted directed graph, such that for every node, the total weight of incoming edges summarizes to at most 1.

Formally, denote a probability network by $G = (V, E)$, where $V$ is the set of $n$ nodes, and $E$ is the set of $m$ directed edges. (See Fig. 1 for an example.) For each edge $(v, u) \in E$ (the solid lines in Fig. 1), it directs from node $v$ to node $u$, and has a weight $w_{v,u} \in [0,1]$. Denote the outgoing neighborhood of node $v$ by $\Gamma_v := \{u | (v, u) \in E\}$, and then it holds that $\sum_{u \in \Gamma_v} w_{v,u} \leq 1$.

With a probability network $G$, the transform matrix $M$ with respect to $G$ is defined as an $n \times n$ real matrix, and the $(v, u)$-th entry (i.e. the entry at the cross of node $v$ and the column corresponding to node $u$) takes value $w_{v,u}$. If there is no edge between node $v$ and node $u$, then the $(v, u)$-th entry takes value 0. As we mentioned above, the entries in any row sum up to at most 1.

**FIGURE 1.** Probability network.

**FIGURE 2.** Random walk.

**B. RANDOM WALK**

Consider a random walk on a given probability network $G = (V, E)$. The walk starts with an initial state (i.e. the state at step 0) $s_0$, whose value is a node from $V$. The walk’s state at step $t > 0$ is selected from the neighborhood of the previous state $s_{t-1}$. For neighbor $u \in \Gamma_{s_{t-1}}$, the probability that it is selected as the value of $s_t$ is $w_{s_{t-1}, u}$. If the total weight of the outgoing edges is strictly smaller than 1, then the walk will stay at where it is (i.e. $s_t = s_{t-1}$) with probability $1 - \sum_{u \in \Gamma_{s_{t-1}}} w_{s_{t-1}, u}$.

For a subset $S \subseteq V$, we say that the walk hits $S$ by step $t$ if and only if there exists $t' \leq t$ such that $s_{t'} \in S$.

**C. THE PROBLEM OF OPTIMIZING THE CONSTRAINED ESTIMATE**

In a problem of optimizing the constrained estimate, a probability network $G = (V, E)$, a step number $t \in \mathbb{Z}^+$, and a constraint (budget) $k \in \mathbb{Z}^+$ are given. Typically, $k \leq n$.

It is required to find a subset $S \subseteq V$ with $|S| \leq k$, such that for a random walk of which the initial state is picked
from $V$ uniformly at random, the probability that it hits $S$ by step $t$ is maximized. For any candidate of such subset $S$, it is called an estimate with constraint $k$.

Consider the example in Fig. 2, in which the dotted lines denote a two-step random walk on the probability network given in Fig. 1. With an initial state $s_0 = A$, to hit node $D$ by step 2, the walk has to jump to node $C$ at step 1, and then jump to node $D$ at step 2. Note that this is the only way to hit node $D$ in no more than two steps, and hence the probability that the walk started from node $A$ hits node $D$ by step 2 is $0.4 \times 0.6 = 0.24$.

For a random walk with given initial state $v$, we denote the event that the walk hits node $u$ by step $t$ as $\mathcal{H}_t^v(u)$. Given a subset $S \subseteq V$, we define $\mathcal{H}_t^v(S) := \cup_{u \in S} \mathcal{H}_t^v(u)$.

### III. THE HARDNESS PROOF

In this section, we will reduce the problem of optimizing the constrained estimate to the problem of vertex cover, and hence the later one is NP-Hard.

Consider an instance of the problem of vertex cover with a given undirected graph $G$. Without any loss of generality, we assume that $S$ does not contain any isolate node of $G$. We construct a probability network $G^*$ with following steps:

1. add all nodes that are not isolated into graph $G$ to graph $G^*$;
2. for each undirected edge $\{v, u\}$ in graph $G$, add two directed edges $(v, u)$ and $(u, v)$ into graph $G^*$;
3. for each node in $G^*$, distribute weight $1$ uniformly to all its outgoing neighbors. That is, for any outgoing neighbor $u$ of a node $v$ in $G^*$, it has weight $w_{v,u} = 1/|r_v|$.

If there is a vertex cover of size $k$ for graph $G$, it is clear that the resulting cover set $S$ is also the optimum estimate of constraint $k$ by step 1 for the probability network $G^*$. In details, for any node $v$ in $G^*$, consider the walk started with $s_0 = v$. If $v \in S$, then $s_0 \in S$. Otherwise, all the outgoing neighbors of node $v$ must belong to $S$ as is a vertex cover, which implies $s_1 \in S$. At all, the walk started with initial state $v$ hits $S$ by step 1 with probability 1. Consequently, we know for a random walk of which the initial state is picked uniformly at random, the probability that it hits $S$ by step 1 is 1, which achieves the maximum.

On the other hand, if there exists subset $S$ of size $k$, such that for a random walk of which the initial state is picked uniformly at random, the probability that it hits $S$ by step 1 is 1, then $S$ is also a vertex cover of size $k$ of graph $G$. To see the reason, let’s assume that there is an edge $(v, u)$ in $G$ with $v \notin S$ and $u \notin S$. In such case, for the walk started with initial state $v$ (or $u$), the probability that it hits $S$ by step 1 is at most $1 - 1/|r_v|$, which is strictly smaller than 1. Thus, a contradiction causes.

As the summary of this section, we have the following Theorem 1.

**Theorem 1.** The problem of optimizing the constrained estimate on a given probability network is NP-Hard.

### IV. THE APPROXIMATION TO THE OPTIMUM

In the last section, we have proved the difficulty of optimizing the constrained estimate. In this section, we will propose an approach which approximates the maximum constrained estimate with a constant ratio.

With a given probability network $G = (V, E)$, for a random walk of which the initial state is picked from $V$ uniformly at random, denote the probability that it hits the subset $S$ by step $t$ as $P_G^t(S)$. Then it holds that

$$P_G^t(S) = \sum_{v \in V} \frac{1}{n} \cdot P(\mathcal{H}_t^v(S))$$

where $P(\cdot)$ denotes the probability that an event happens.

As a function on the subsets of $V$, $P_G^t(\cdot)$ has two important properties: monotonicity and submodularity, which are formally described in the following lemmas.

**Lemma 1 (Monotonicity).** Function $P_G^t(\cdot)$ is monotone on the subsets of $V$. That is, for any subsets $S \subseteq V$ and $S' \subseteq V$, if $S \subseteq S'$, then it is implied that $P_G^t(S) \leq P_G^t(S')$.

**Proof.** Note that for any initial state $v \in V$, $\mathcal{H}_t^v(S) \subseteq \mathcal{H}_t^v(S')$, which implies that $P(\mathcal{H}_t^v(\cdot))$ is monotone on the subsets of $V$. Thus, the monotonicity of $P_G^t(\cdot)$ follows.

**Lemma 2 (Submodularity).** Function $P_G^t(\cdot)$ is submodular on the subsets of $V$. That is, for any subsets $S \subseteq V$ and $S' \subseteq V$, if $S \subseteq S'$, then it is implied that for any node $u \in V$, $P_G^t(S \cup \{u\}) - P_G^t(S) \leq P_G^t(S' \cup \{u\}) - P_G^t(S')$.

**Proof.** It is sufficient to prove that for any initial state $v \in V$, $P(\mathcal{H}_t^v(\cdot))$ is submodular on the subsets of $V$. Note the fact that

$$P(\mathcal{H}_t^v(S \cup \{u\})) = P(\mathcal{H}_t^v(S) \cup \mathcal{H}_t^v(u))$$

is submodular on the subsets of $V$. Note the fact that

$$P(\mathcal{H}_t^v(S \cup \{u\})) = P(\mathcal{H}_t^v(S)) + P(\mathcal{H}_t^v(u)) + P(\mathcal{H}_t^v(S) \cap \mathcal{H}_t^v(u))$$

Then it follows that

$$P(\mathcal{H}_t^v(S \cup \{u\})) = P(\mathcal{H}_t^v(S'))$$

where the inequality holds since $\mathcal{H}_t^v(S) \subseteq \mathcal{H}_t^v(S')$ implies $P(\mathcal{H}_t^v(S) \cap \mathcal{H}_t^v(u)) \leq P(\mathcal{H}_t^v(S') \cap \mathcal{H}_t^v(u))$.

This finishes the proof.

Nemhauser and Fisher et al. have shown that the standard greedy algorithm (also known as hill climbing) can achieve the $(1 - 1/e)$-approximation\(^1\) to the optimum of any monotone and submodular function [4], [8]. Thus, we propose

\(^1\)Here, $e$ stands for the base of the natural logarithm.
the following algorithm to approximate the optimum of the constrained estimate (Algorithm 1)\(^2\).

Algorithm 1: Hill climbing for the \((1 - 1/e)-\)approximation of optimum estimate of constraint \(k\) by step \(t\), on the given probability network \(G = (V, E)\).

\[
S := \emptyset \\
\text{while } |S| < k \text{ do} \\
\quad u^* := \arg\max_u \mathcal{P}_G(S \cup u) \\
\quad S = S \cup \{u^*\} \\
\text{end while}
\]

Assuming \(\mathcal{P}_G(S \cup u)\) can be calculated in complexity \(C\), the complexity of Algorithm 1 is \(O(k \cdot n \cdot C)\), which is \(O(n^2 \cdot C)\) as \(k \leq n\). (Recall that \(n\) is the number of nodes in the network.) In the remaining argument, we will show that \(C = O(n^{2.3737} \cdot \log t)\).

At first, for any subset \(S \subseteq V\), it holds that
\[
\mathcal{P}_G(S) = \sum_{v \in V} \frac{1}{n} \cdot \mathbb{P}(\mathcal{H}_v^t(S)) = \sum_{v \in V} \frac{1}{n} \cdot (1 - \mathbb{P}(\mathcal{H}_v^t(V) \setminus \mathcal{H}_v^t(S))) = 1 - \sum_{v \in V \setminus S} \frac{1}{n} \cdot \mathbb{P}(\mathcal{H}_v^t(V) \setminus \mathcal{H}_v^t(S))
\]

The last equality holds because event \(\mathcal{H}_v^t(V) \setminus \mathcal{H}_v^t(S)\) means the walk started at node \(v\) doesn’t hit any node in \(S\) by step \(t\), and consequently, it follows that \(\mathcal{H}_v^t(V) \setminus \mathcal{H}_v^t(S) = \emptyset\) with \(v \in S\).

Recall the given probability network \(G\), we construct a special transform matrix \(M_S\) with respect to the subset \(S\). Consider the \((v, u)\)-th entry of \(M_S\). If \(u \notin S\), the entry has value \(w_{v,u}\), which is the probability that the walk starting at node \(v\) selects node \(u\) in the coming step. Otherwise, the \((v, u)\)-th entry has value 0. Thus, we have
\[
\sum_{v \in V \setminus S} \frac{1}{n} \cdot \mathbb{P}(\mathcal{H}_v^t(V) \setminus \mathcal{H}_v^t(S)) = \frac{1}{n} \cdot \mathbf{e}_{V \setminus S}^T \cdot M^t \cdot \mathbf{e},
\]

where \(\mathbf{e}\) is the all-one column vector and \(\mathbf{e}_{V \setminus S}\) is the transpose of column vector \(\mathbf{e}_{V \setminus S}\), which has \(n\) boolean entries for the nodes in \(V\) and only the entries corresponding to the nodes outside \(S\) take value 1. Note that \(M^t\) can be calculated in at most \(\log t\) layers of recursions and in each layer, it costs at most \(O(n^{2.3737})\) for the matrix multiplication. Thus, we have \(C = O(n^{2.3737} \log t)\), which shows that the fact of a \((1 - 1/e)-\)approximation of the optimum constrained estimate can be found in polynomial time.

\(^2\)In theoretical computer science, approximation algorithms are efficient algorithms that find approximate solutions to NP-hard optimization problems with provable guarantees on the distance of the returned solution to the optimal one. The guarantee of such algorithms is a multiplicative one expressed as an approximation ratio, i.e., the optimal solution is always guaranteed to be within a (predetermined) multiplicative factor of the returned solution.

V. COUPLING TO THE INFLUENCE CASCADING

In this section, we show that the random walk can be coupled with an influence cascading process on the same probability network. As a consequence, this coupling relation implies an alternative way to calculate \(\mathcal{P}_G(S)\).

With a given probability network \(G = (V, E)\), each node \(v \in V\) is associated with a state of the binary value: active, or inactive.

The influence cascading starts with a subset \(S \subseteq V\), called the distributors. The distributors are always active. Given a subset \(S\) of distributors, define the random variables \(x^t_S(v)\) for all step \(t\) and node \(v\), such that \(x^t_S(v) = 1\) if node \(v\) is active at step \(t\); and \(x^t_S(v) = 0\) otherwise.

At step 0, only nodes inside \(S\) are active. At step \(t\), for any node \(v \in V \setminus S\), it becomes active with probability \(\sum_{w \in \Gamma_v} w_{v,u} \cdot x^t_{S-1}(u)\).

Now we are ready to formally introduce the coupling relation between the random walk and the influence cascading (Theorem 2). It should be mentioned that a similar relation was revealed between the random walks and the voter model [5], and a quantitative conclusion was proved on the undirected graphs [3].

**Theorem 2.** \(\mathbb{P}(\mathcal{H}_v^t(S)) = \mathbb{P}(x^t_S(v) = 1)\).

**Proof.** Note that for any step \(t\), if \(v \in S\), then it always holds that \(\mathbb{P}(\mathcal{H}_v^t(S)) = 1 = \mathbb{P}(x^t_S(v) = 1)\). Thus, we only need to consider the cases when \(v \notin S\). We prove the conclusion by the induction on \(t\). For \(t = 0\), as long as \(v \notin S\), we have \(\mathbb{P}(\mathcal{H}_v^t(S)) = 0 = \mathbb{P}(x^t_S(v) = 1)\).

For \(t > 0\), assume that \(\mathbb{P}(\mathcal{H}_v^t(S)) = \mathbb{P}(x^t_S(v) = 1)\) holds for all \(t' < t\). With \(v \notin S\), we have
\[
\mathbb{P}(x^t_S(v) = 1) = \mathbb{E}(x^t_S(v)) = \mathbb{E}(\sum_{u \in \Gamma_v} w_{v,u} \cdot x^{t-1}_S(u)) = \sum_{u \in \Gamma_v} w_{v,u} \cdot \mathbb{E}(x^{t-1}_S(u)) = \sum_{u \in \Gamma_v} w_{v,u} \cdot \mathbb{P}(x^{t-1}_S(u) = 1) = \sum_{u \in \Gamma_v} w_{v,u} \cdot \mathbb{P}(\mathcal{H}^{t-1}_v(S)) = \mathbb{P}(\mathcal{H}^t_v(S)),
\]

where \(\mathbb{E}(\cdot)\) denotes the expectation of the random variable. This finishes the proof.

With the given network \(G = (V, E)\) and a subset \(S\) of distributors, we define a special transform matrix \(F_S\) of dimension \(n \times n\). For the \((u, v)\)-th entry of \(F_S\), it takes value \(w_{v,u}\) if \(v \notin S\); otherwise, it takes value 1 if \(u = v\) and 0 if \(u \neq v\). Thus we have
\[
\mathbb{P}(x^t_S(v) = 1) = \mathbf{e}_v^T \cdot F_S \cdot \mathbf{e}_v
\]

where \(\mathbf{e}_v\) is the character vector of node \(v\) (i.e. entry corresponding to \(v\) takes value 1; other entries take value 0), and \(\mathbf{e}_S\) is the transpose of column vector \(\mathbf{e}_S\), in which only
the entries corresponding to the nodes inside $S$ take value 1 (other entries take value 0). This result also shows how to calculate $\mathbb{P}(H^t_i(S))$, and consequently gives an alternative way to calculate $P^t_G(\cdot)$.  

VI. DISCUSSIONS

A. THE STRICT ESTIMATE

In the previous sections, we studied the estimate of the random walk by step $t$. The whole thing becomes much easier if we only want to estimate the situation exactly at step $t$. Formally speaking, given the probability network $G = (V, E)$, let $R^t_v(S)$ denote the event that with the random walk initialized at node $v$ hits subset $S \subseteq V$ at step $t$, i.e. $s_t \in S$. Then with a size budget $k$, how to find a subset $S$ of size $k$ such that for the random walk of which the initial state $v$ is picked from $V$ uniformly at random, the probability $\mathbb{P}(R^t_v(S))$ is maximized?

This problem is actually easy to solve. Let $M$ denote the transform matrix with respect to the given probability network $G$. Then we have

$$\sum_{v \in V} \frac{1}{n} \cdot \mathbb{P}(R^t_v(S)) = \frac{1}{n} \cdot e' \cdot M^t \cdot e_S$$

where $e'$ is the transpose of the all-one column vector and $e_S$ is the character vector with respect to subset $S$ (entries corresponding to nodes in $S$ take value 1; other entries take value 0). Hence, to find the strict estimate at step $t$, we only need to calculate the vector $e' \cdot M^t$ and select the nodes that corresponding to the top $k$ entries to form the subset $S$.

B. THE BIASED INITIALIZER

We have been considering the random walk when the initial state is picked uniformly at random. However, this restriction can be much flexible. It is easy to find out that if the initial state is determined according to some specific distribution on all nodes, the parameters in this document are still valid after replacing the uniform probability $1/n$ with the distribution probability $\mathbb{P}(s_0 = v)$. This fact also implies that the initial state can be restricted over a subset of all the nodes.

In general, we can use an $n$-dimensional vector $s_0$ as the initial state of the random walk. As long as the probability with which the neighbor of $s_0$ is selected is a real number between 0 and 1, and all their sum is 1, the problem of optimizing the constrained estimate defined in this paper can be generalized to this case.

REFERENCES


