Maximizing Influence-Based Group Shapley Centrality

Extended Abstract

Ruben Becker
Gran Sasso Science Institute
L’Aquila, 67100, Italy
ruben.becker@gssi.it

Gianlorenzo D’Angelo
Gran Sasso Science Institute
L’Aquila, 67100, Italy
gianlorenzo.dangelo@gssi.it

Hugo Gilbert
Université Paris-Dauphine, Université
PSL, CNRS, LAMSADE
75016 Paris, France
hugo.gilbert@dauphine.psl.eu

ABSTRACT

A key problem in network analysis is the influence maximization problem, which consists of finding a set $S$ of at most $k$ seed users in a social network, such that the spread of information from $S$ is maximized. We investigate the problem of choosing the best set of seeds when there exists an unknown pre-existing set of seed nodes. Our work extends the one of Chen and Teng (WWW’17) who introduced the so-called Shapley centrality of a node to measure the efficiency of nodes acting as seeds within a pre-existing but unknown set of seeds. We instead consider the question: Which set of cardinality $k$ to target in this kind of scenario? The resulting optimization problem reveals very challenging, that is, assuming common computational complexity conjectures, we obtain strong hardness of approximation results. Nevertheless, we design a greedy algorithm which achieves an approximation factor of $\frac{1-1/e}{k} - \epsilon$ for any $\epsilon > 0$, showing that not all is lost in settings where $k$ is bounded.

KEYWORDS

Influence Maximization; Group-Shapley Value

ACM Reference Format:

1 MOTIVATION

Node centrality and spread of information or influence are two main topics in network analysis. The former regards the problem of determining the most important nodes in a network according to some measure of importance, while the latter studies mathematical models to represent how information spreads in a Social Network (SN) in which nodes are able to communicate with each other.

In order to measure the centrality of nodes in a network a real-valued function, called centrality index, associates a real number with each node that reflects its importance or criticality within the network. Chen and Teng [4] initiated the study of the interplay between spreading dynamics and network centrality by defining two centrality indices based on dynamic models for influence diffusion: the single node influence centrality, which measures the centrality of a node by its capability of spreading information when acting alone, and the Shapley centrality, which uses the Shapley value to measure the capability of a node to increase the spreading capacity of a pre-existing but unknown group of nodes. The Shapley value – a well-known concept from cooperative game theory – models this scenario by assessing the expected relevance of each player within a random subset of players (also called coalition), where the expectation is taken over the possible coalitions. More formally, given a characteristic function $\tau$ that maps each coalition to the total payoff that that coalition receives, the Shapley value of a player $i$ is defined as the expected payoff that $i$ adds to any coalition, w.r.t. the function $\tau$. The Shapley centrality index studied by Chen and Teng [4] measures the centrality of a node by using the Shapley value and the spreading function $\sigma$ as characteristic function.

Most centrality indices neglect the relevance that coalitions of individuals and their coordination play in SNs. For this reason, many centrality indices have been generalized to group centrality indices which are real-valued functions over subsets of nodes instead of single nodes. A group centrality index is fundamentally different from a combination of the individual centrality indices of the nodes in the group, as it captures the relevance of the set as a whole. This work extends Chen and Teng’s notion of influence-based Shapley centrality from single nodes to groups of nodes by using the concept of the Group Shapley value. Our Influence-based Group Shapley (IGS) centrality associates to a set $S$ of nodes, the expected gain in influence that $S$ adds to a random pre-existing seed set $T$. We investigate the problem of finding a set $S$ of cardinality $k$ with highest IGS value. We believe that this way of evaluating the importance of a seed set is of high practical interest. Assume an entity wants to spread a piece of news, while having a budget to influence $k$ users, at the same time knowing that already some users are aware of the information and will spread it anyhow. The central entity, however, may have no knowledge about who these users are. In this case, it should target a set of seed users with large IGS value.1

2 PRELIMINARIES

The Group Shapley Value. In cooperative game theory [3, 10], a game on $n \geq 2$ players is commonly formalized by a characteristic function $\tau : 2^n \to \mathbb{R}$ that assigns to every subset $S \subseteq [n]$ of players, also called a coalition, a value $\tau(S)$. The Group Shapley value $[5, 9]$, for a subset $S \subseteq [n]$ of players in a game $\tau$ is defined as

$$\phi^S_{\tau}(S) := \sum_{T \subseteq [n] \setminus S} \frac{|T|!(n - |S| - |T|)!}{(n - |S| + 1)!} \cdot (\tau(T \cup S) - \tau(T)),$$

i.e., the Group Shapley value quantifies how much $S$ is able to add in the game $\tau$ to a random subset $T$ that is generated as follows: We first sample an integer $t \in \{0, \ldots, n - |S|\}$ uniformly at random and then pick a set $T$ of size $t$ in $[n] \setminus S$ uniformly at random.

1See [1] for the complete proofs of all results reported on in this extended abstract.
Influence Maximization (IM). We are interested in the Group Shapley value for functions that describe information propagation in SNs. Two of the most popular models for describing such information propagation are the Independent Cascade (IC) and Linear Threshold (LT) models [7]. In both models, we are given a directed graph \( G = (V, E) \) where \( V \) is a set of \( n \) nodes, values \( \{ p_{uv} \in [0, 1] : (u, v) \in E \} \) and an initial node set \( A \subseteq V \) called seed set. A spread of influence from \( A \) is defined as a randomly generated sequence of node sets \( (A_1, \ldots, A_{t^*}) \) called \( \sigma \). Furthermore, for all settings where \( k \) is an integer, the Max-Shapley-Group Hypothesis holds [8]. We provide a reduction which yields the same result.

We are interested in the Group Shapley value of nodes w.r.t. the influence spread at time \( t \). For a set \( (A_1, \ldots, A_{t^*}) \), we call this the Influence-based Group Shapley value \( \phi_{Sh}(A) \), i.e., the Group Shapley value under a cardinality constraint. Both the IC and LT model are special cases of the more general Triggering Model, see [7, Proofs of Theorem 4.5 and 4.6] that we consider in this work.

Influence-based Group Shapley Centrality. Chen and Teng [4] consider the Shapley value of nodes w.r.t. the influence spread function \( \sigma \) in a SN modeled by the Triggering Model. They use the resulting Shapley centrality \( \phi_{Sh}(i) \) as a measure of centrality of node \( i \). In this work, we consider \( \phi_{Sh} \), i.e., the Group Shapley value w.r.t. \( \sigma \). We call this value the Influence-based Group Shapley (IGS) centrality of \( S \). We refer to it as \( \phi_{Sh}(S) \) omitting \( \sigma \) as an index.

The central problem of this paper consists in finding sets \( S \) of size at most \( k \) that have large IGS centrality among all such sets. We call this the Max-Shapley-Group problem. The most important tool for studying this problem are so-called RR sets [2, 12]. In fact, there exists a concise formulation of \( \phi_{Sh}(S) \) in terms of RR sets:

**Lemma 2.1** (IGS centrality via RR sets). For any \( S \subseteq V \),

\[
\phi_{Sh}(S) = n \cdot \mathbb{E}_p \left[ \frac{1_{R_t \cap S \neq \emptyset}}{|R_t \cap S| + 1} \right].
\]

**3 HARDNESS OF APPROXIMATION**

Unfortunately, Max-Shapley-Group in the IC model is up to a constant factor, as hard to approximate as Densest-k-Subgraph.

**Theorem 3.1.** Let \( \alpha \in (0, 1) \). If there is an \( \alpha \)-approximation algorithm for Max-Shapley-Group, then there is an \( 1/\alpha \)-approximation algorithm for Densest-k-Subgraph.

A number of strong hardness of approximation results are known for Densest-k-Subgraph: (1) Densest-k-Subgraph cannot be approximated within \( 1/n^{o(1)} \) if the Gap Exponential Time Hypothesis holds [8]. (2) Densest-k-Subgraph cannot be approximated within any constant if the Unique Games with Small Set Expansion conjecture holds [11]. (3) Densest-k-Subgraph cannot be approximated within \( n^{-O(\log \log n)} \) for some constant \( c \) if the Exponential Time Hypothesis holds [8]. We provide a reduction which yields the same hardness results for Max-Shapley-Group. In particular, according to (1) and our reduction, it is unlikely to find anything better than an \( (n^{-c}) \)-approximation for Max-Shapley-Group, where \( c \) is a constant. Furthermore, for all settings where \( k = O(n^c) \), such an algorithm is implied by our result in Section 4.

**4 A SIMPLE APPROXIMATION ALGORITHM**

A good approximation result, as for example a constant-factor approximation, is unlikely for the Max-Shapley-Group problem. We however obtain a positive result for small values of \( k \).

Approximating \( \phi_{Sh} \) through RR sets. By sampling a sufficient number of RR sets, we can give a set function \( \phi_{Sh} \) that with high probability approximates \( \phi_{Sh} \) to within a factor of \( 1 \pm \epsilon \). The idea is to approximate the expected value in Lemma 2.1 via a Chernoff bound. This is captured by the following lemma.

**Lemma 4.1.** Let \( \epsilon \in (0, 1) \) and let \( R_1, \ldots, R_t \) be a sequence of \( t \) RR sets. For a value of \( t \) that is polynomial in \( n \) and \( \epsilon^{-1} \), w.h.p.,

\[
\phi_{Sh}(S) = \frac{n}{t} \sum_{i=1}^t \frac{1_{R_i \cap S \neq \emptyset}}{|R_i \cap S| + 1}
\]

is a \((1 \pm \epsilon)\)-approximation of \( \phi_{Sh} \).

The Harmonic-Max-Hitting-Set problem. Lemma 4.1 suggests to sample a near-linear number \( t \) of RR sets and compute a set of nodes \( S \) that maximizes \( \phi_{Sh}(S) \). We call the resulting problem the Harmonic-Max-Hitting-Set problem: the input consists of a set \( X = \{x_1, \ldots, x_n\} \), a set \( Z = \{Z_1, \ldots, Z_m\} \) of subsets of \( X \), and an integer \( k \). The task is to find a subset \( S \subseteq X \) s.t. \( |S| \leq k \) maximizing

\[
f_{Z}(S) = \sum_{i=1}^m \frac{1_{Z_i \cap S \neq \emptyset}}{|Z_i \cap S|}.\]

This is a non-linear variant of the well-known Max-Hitting-Set problem (itself equivalent to the Max-Set-Cover problem [6]) in which the objective function is \( \sum_{i=1}^m 1_{Z_i \cap S \neq \emptyset} \). The problem of maximizing \( \phi_{Sh} \) can be stated as a Harmonic-Max-Hitting-Set problem by letting \( X = V \) and \( Z \) be the set of generated RR sets.

**Approximation Algorithm.** Consider an instance \((X, Z, k)\) and define the following set function \( h_{Z}(S) = \sum_{i=1}^m 1_{Z_i \cap S \neq \emptyset}/|Z_i| \). Note the similarity between \( h_{Z} \) and \( f_{Z} \). The approximation algorithm that we propose is to greedily maximize \( h_{Z} \) instead of \( f_{Z} \). Why would this be a good idea? (1) The set function \( h_{Z} \) is monotone and submodular; thus the greedy algorithm will yield a \( 1 - 1/e \) approximation to maximizing \( h_{Z} \). (2) Given a set \( S \subseteq X \) with \( |S| \leq k \), it holds that \( f_{Z}(S) / h_{Z}(S) \geq f_{Z}(S)/h_{Z}(S)/k \), that is, the error when passing from \( h_{Z} \) to \( f_{Z} \) is at most \( k \). Hence, if we denote by \( S^*_h \) (resp. \( S^*_f \)) an optimal solution of size \( k \) for maximizing \( f_{Z} \) (resp. \( h_{Z} \)), we have that \( h_{Z}(S^*_h) \geq h_{Z}(S^*_f) \geq f_{Z}(S^*_h)/k \). Now let \( S \) be the solution of size \( k \) returned by the greedy algorithm. Then, \( S \) is a \((1 - 1/e)/k \) approximation to maximizing \( f_{Z} \) as

\[
f_{Z}(S) \geq h_{Z}(S) \geq \left( 1 - \frac{1}{e} \right) \cdot h_{Z}(S^*_h) \geq \left( 1 - \frac{1}{e} \right) \cdot f_{Z}(S^*_f)/k.
\]

**Theorem 4.2.** Let \( \epsilon \in (0, 1) \). There is an algorithm with running time polynomial in \( n \) and \( \epsilon^{-1} \) that, with high probability, returns a \((1-1/e)/k - \epsilon\)-approximation for the Max-Shapley-Group problem.

**ACKNOWLEDGMENTS**

This work was partially supported by the Italian MIUR PRIN 2017 Project “ALGADIMAR” Algorithms, Games, and Digital Markets.
REFERENCES